In compliance with the
Canadian Privacy Legislation
some supporting forms
may have been removed from
this dissertation.

While these forms may be included
in the document page count,
their removal does not represent
any loss of content from the dissertation.

University of Alberta

Prediction of Protein Chemical Shifts From Atomic Coordinates

by

Stephen Neal

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science in Pharmaceutical Sciences

Faculty of Pharmacy and Pharmaceutical Sciences

Edmonton, Alberta
Fall 2003

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisisitons et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-87846-5
Our file Notre référence
ISBN: 0-612-87846-5

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

Canadä

University of Alberta

Library Release Form

Name of Author: *Stephen Neal*

Title of Thesis: *Prediction of Protein Chemical Shifts From Atomic Coordinates*

Degree: *Master of Science*

Year this Degree Granted: *2003*

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Aug 29, 2003

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Prediction of Protein Chemical Shifts from Atomic Coordinates** submitted by **Stephen John Neal** in partial fulfillment of the requirements for the degree of **Master of Science. in Pharmaceutical Sciences.**

Dr. D. Wishart (Supervisor)

Dr. S. McQuarrie (Chair)

Dr. G. Kotovych

Date: AUG 29, 2003

Abstract

The relationship between protein structure and chemical shifts is ultimately determined at the level of quantum mechanics. However, it is also possible to determine chemical shifts using statistical methods in combination with classical or semi-classical models of these quantum phenomena. Two computer programs were created to explore and apply these relationships. One program, MINER, applied the techniques of data mining and machine learning to tabulate useful structure/shift connections. Using a database of proteins for which high resolution X-ray protein structures and chemical shifts were available, MINER generated tables that enabled prediction of chemical shifts directly from polypeptide structure. The other program, SHIFTX, took protein coordinate data as input and produced chemical shift predictions for all NMR-detectable nuclei in the protein. These predictions were made using the tables from MINER as well as classical and semi-classical formulas for ring current, electrostatic, and hydrogen bonding effects. SHIFTX quickly and accurately predicts chemical shifts for a wide range of diamagnetic protein structures.

"Abide me, if thou dar'st; for well I wot
Thou runn'st before me, shifting every place,
And dar'st not stand, nor look me in the face.
Where art thou now?"

-- A somewhat consternated Demetrius
William Shakespeare's
A Midsummer-Night's Dream
Act III, Scene II

Acknowledgements

The author gratefully acknowledges the contributions of the many people who have contributed to this work in one way or another. Dr. David Wishart patiently shepherded this "two year project" through four and a half years, contributing his vast knowledge and experience at every step. Alex Nip, Haiyan Zhang, and Wei Xia also generously shared their time and expertise. Dr. Steve McQuarrie and Dr. George Kotovych sat on my examining committee, read this monster, provided valuable feedback, and most importantly didn't ask me about anisotropy or any of that other nuclear magnetic whatsit. Joyce Johnson of the Faculty of Pharmacy dealt cheerfully with all the problems caused by a troublesome out-of-country graduate student who just couldn't do things the way everybody else did. Dr. Charles Bourassa shared his keen insight into how academia and research *actually* function, in contrast to the recruiting brochure version. My sister, Dr. Joanne Neal, proofread *far* above and beyond the call of duty, and my niece Dakota generously loaned me her mom to do all that editing. My *other* sister Sally Neal provided aesthetic sense and moral support, to say nothing of the last-possible-moment duplication-and-distribution blitz. My parents, Jack and Joan Neal, made contributions beyond measure. Those hardened veterans of the thesis wars, Diane Cantwell, Jessica Murphey, and Shannon Williams, shared war stories, inspiration, and sympathy, and tolerated my outbursts of sound and fury; Shannon also made the invaluable contribution of the proper nomenclature for such an important document as this. Michael Kaplan, the Adobe Atmosphere Zombies, and Adobe Systems Inc. provided indulgences of time (to say nothing of funding!) that made the completion of this thesis even remotely plausible. Roger Jorgenson provided films, coffee, and conversation, and helped prove my secondary hypothesis: that old guys can go back to school.

Thanks, everybody - here's to ya!

Table of Contents

Chapter 1: Introduction 1

Preface 1

Background 4

Proteins 4

Protein Structures 8

Hydrogen Bonds 10

Disulfide Bonds 12

The Significance of Protein Structure 12

Protein Structure Determination 13

Chemical Equivalence 16

Chemical Shifts 17

Diamagnetic Effects 19

Paramagnetic Shielding (Local) 19

Anisostropic Shielding 20

Ring Currents 20

Electrostatic Effects 20

Solvent Effects 21

Theory and Mathematical Prediction of Chemical Shifts 21

Random Coil Shifts 22

Databases 24

Atomic Nomenclature in Proteins 25

Structure Resolution **26**

The BioMagResBank **27**

Outline and Objectives of this Dissertation 27

References 28

Chapter 2: MINER 30

Introduction 30

Methods 31

Analytical Approach **33**

Tables **36**

The Best Guess **37**

Choosing Axes **37**

Building Tables **39**

Splines **40**

Making Predictions **42**

Evaluation **44**

Picking The Winner **45**

Numerical Optimization **47**

Testing **52**

Discussion 54

Conclusion 56

References 57

Chapter 3: SHIFTX 58

Introduction 58

Methods 60

File Input **61**

Hydrogen Placement **62**

Ring Current Effects **62**

Electric Field Effects **65**

Hydrogen Bond Effects **66**

Empirical Chemical Shift Hypersurfaces **69**

Training and Testing Databases 73

Results and Discussion 76

Hypersurfaces **79**

Performance of ^{1}H Predictions **84**

Performance of ^{13}C and ^{15}N Predictions **85**

Secondary Validation **90**

Applications **92**

Limitations **99**

Availability **102**

References 104

Chapter 4: Conclusion 107

References 112

List of Tables

Table 1-1: The twenty amino acids and their abbreviations. 5

Table 1-2: Random coil shifts. 23

Table 1-3: Sidechain protons 26

Table 2-1: Example Table 37

Table 2-2: Example MINER table.. 39

Table 2-3: Contrived data set with 4 data points and two factors. 48

Table 2-4: Processing the data set of Table 2-3. 48

Table 2-5: Second round of processing the data set of Table 2-3. 49

Table 2-6: Test cases for MINER.. 53

Table 3-1: Residue types containing aromatic rings 63

Table 3-2: Physical or Derived Property Parameters Used By MINER. 70

Table 3-3: PDB and BMRB files used to calibrate the backbone shift predictions 74

Table 3-4: PDB and BMRB files used to calibrate the side chain H shift predictions. ... 75

Table 3-5: Correlation between the observed and predicted backbone chemical shifts... 77

Table 3-6: Correlation between the observed and predicted side chain chemical shifts .. 78

Table 3-7: Relative influences of various factors on secondary shifts. 80

Table 3-8: Principal component analysis of the prediction factors for lysine HA shifts.. 83

Table 3-9: Proteins used to validate SHIFTX performance on previously unseen data... 91

Table 3-10: Results for ten proteins not used in the training data 91

Table 3-11. Comparison between initial and final assignments for Mt0807 95

List of Figures

Figure 1-1: Generalized Amino Acid 7

Figure 1-2: Protein Backbone 7

Figure 1-3: Torsion Angles 9

Figure 1-4: Ramachandran plot 10

Figure 1-5: Atomic Bonding Environments. 18

Figure 1-6: Records from a PDB file. 24

Figure 2-1: Comparing factors applied together to factors applied singly. 46

Figure 3-1: An example hypersurface 71

Figure 3-2: Scatterplots comparing observed vs. predicted shifts for backbone nuclei ... 88

Figure 3-3: Plot of the observed vs. predicted $^{13}C\alpha$ shift of ribonuclease H 93

Figure 3-4: Scatterplots of the resolution vs. correlation for the backbone nuclei 97

Figure 3-5: A screen shot of the SHIFTX web server. 103

List of Symbols and Abbreviations

Å: Angstrom = 10^{-10} meters

BPTI: Bovine pancreatic trypsin inhibitor

CA: Alpha carbon

CB: Beta carbon

CO: Carbonyl carbon

EM: Electromagnetic

HA: Alpha hydrogen

HN: Amide hydrogen (on a protein backbone). Often labeled as simply "H".

HNCACB: Amide Hydrogen/Alpha Carbon/Beta Carbon

HSQC: heteronuclear single quantum correlation

N15: Amide nitrogen (on a protein backbone)

NMR: nuclear magnetic resonance

NOE: Nuclear Overhauser effect

NOESY: Nuclear Overhauser effect spectroscopy

PDB: Protein Data Bank

RF: Radio frequency

RMSD: Root Mean Squared Deviation

TOCSY: total correlation spectroscopy

Chapter 1: Introduction

Preface

Proteins account for nearly 95% of all known drug targets and represent a rapidly growing percentage of prescription pharmaceuticals (Dean, Zanders and Bailey, 2001; Maggio and Ramnarayan, 2001). Given their importance in human health and disease, it is little wonder that increasingly more money and effort is going in to improving our understanding of protein structure, function, and activity. However, unlike small molecule drugs or vitamins, proteins are not simple molecules that are easily isolated and analyzed. Consequently, one of the principle challenges in life science research over the past 50 years has been to develop better methods or faster techniques which can simplify or improve the identification, purification, and structure determination (Heinemann, Illing, and Oschkinat, 2001) of peptides and proteins. One protein analysis technique that has grown in importance over the past 25 years is NMR (nuclear magnetic resonance) spectroscopy. In addition to its utility for understanding protein-ligand interactions and protein dynamics, NMR is one of only two methods (the other being X-ray crystallography) which allows scientists to determine the three dimensional structure of proteins to atomic resolution (Creighton, 1993; Lesk, 2001).

Fundamental to every aspect of protein NMR is the assignment or determination of the chemical shifts for each of the atoms found in a given protein molecule. Chemical shifts represent the characteristic NMR absorption frequencies for each NMR active atom in a molecule. Because chemical shifts reflect the cumulative effects from a plethora of atomic and electronic interactions (covalent and non-covalent), they actually contain vital

structural information. As such, chemical shifts can be used to interpret, refine and even predict the three dimensional structure of peptides and proteins (Wishart and Case, 2001).

However, the interpretation and prediction of chemical shifts – particular for proteins – is exceedingly difficult (Heinemann, Illing, and Oschkinat, 2001). Efforts aimed at developing accurate protein chemical shift calculators based on quantum or semi-classical physics have been ongoing for more than 30 years (Case, 2000; Szilagyi, 1995; Williamson and Asakura, 1997). Unfortunately, many of these efforts have failed and, until quite recently, no computer program or method had been developed which could calculate protein chemical shifts with sufficient precision to make them useful to the biomolecular NMR community (Xu and Case, 2002). While "first principle" approaches based on pure quantum physics have been relatively unsuccessful, there are other, less mathematically rigorous approaches which potentially could yield far more useful results. Specifically, there exists a large body of experimental information in the form of known protein chemical shifts (The BioMagResBank -- Seavey et al., 1991) and corresponding known protein structures (Berman et al., 2000). In principle, if there are a sufficient number of observations (shifts and coordinates) in this experimental data, then it should be possible to extract heuristic rules or heuristic formulae that predict chemical shifts from atomic coordinates. The quantity of data or prior (or experimentally observed) knowledge makes the challenge of developing a protein chemical shift "predictor" an ideal task for a branch of computer science called data mining. Data mining is a computationally intensive data analysis technique which employs a variety of statistical and machine learning methods to comb through vast quantities of data and

extract patterns or useful "knowledge" from seemingly uninterpretable collections of text or numbers (Ramakrishnan and Grama, 1999; Mitchell, 1997).

This thesis is concerned with developing novel data mining tools and applying them to the prediction and/or calculation of protein chemical shifts. Specifically, the working hypothesis for this thesis is that *data mining, in concert with well-known physical principles (and corresponding formulae) about chemical shifts can be used to create a computer program that would be capable of accurately and rapidly predicting protein chemical shifts (^{1}H, ^{13}C and ^{15}N) from protein coordinate data.*

By attempting to develop a rapid and accurate chemical shift calculator we believe it could have a significant impact on protein NMR. Specifically, an accurate protein chemical shift calculator could allow protein structures to be better refined or resolved using a process called chemical shift refinement or chemical shift minimization (Oldfield, 1995). Likewise a chemical shift calculator could also assist in the estimation or initial prediction of peptide or protein backbone (and side chain) dihedral angles. Such a program could also be used in the generation and refinement of peptide and protein structures using NOE-independent techniques – such as chemical shift threading or residual dipolar coupling analysis (Wishart and Case, 2001). Furthermore, a chemical shift calculator could be used to facilitate and check protein chemical shift assignments and validate protein chemical shift referencing – a key challenge in much of biomolecular NMR. There are, no doubt, many other applications to protein chemistry and NMR spectroscopy that such a predictive tool could have. However, before delving into the details and potential applications of this computational tool, it is perhaps worthwhile to

take a moment to review some of the key topics of this thesis and to put some of the material already mentioned into a better context.

Background

Proteins

Proteins are manufactured by all living cells, and play key roles in the life, death, and reproduction of those cells. If an analogy is made between a cell and a factory, proteins constitute not only most of the bricks and beams that make up the factory's structures, but also most of the workers and machines of that factory. Each type of protein has evolved to play some role in the cell, and its three-dimensional structure and chemical characteristics reflect that role.

Structurally speaking, proteins are perhaps the most complex chemical entities in nature. Typically consisting of thousands of atoms, linked together and packed together in a multitude of different and complex shapes, proteins are also among the largest molecules known to exist. Proteins are heteropolymers and as such, a protein may be considered as a polymer consisting of a sequence of different monomers (called amino acids or residues) that have been covalently linked into a single continuous unbranched chain. All cells build proteins from a set of twenty amino acids, in the same way that the English language builds words from the twenty-six letters of the alphabet. Each type of protein is defined by its amino acid sequence, which in turn defines its structure. Note that all proteins with the same (or similar) amino acid sequence have the same (or similar) structure and function. The twenty naturally-occurring amino acids are listed in Table 1-1. Included in this list are the codes and abbreviations used to represent them in

the shorthand used to describe protein primary structure, which is just the list of amino acids that make up a protein.

Chemical Name	Three-Letter Code	One-Letter Code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Table 1-1: The twenty amino acids and their abbreviations.

Using this list it is possible to describe the polymeric structure of proteins in a simple sequential list of letters or abbreviations. For instance, the amino acid sequence (order of amino acid residues) for the protein known as BPTI (bovine pancreatic trypsin inhibitor) can be described using the three letter code:

Arg Pro Asp Phe Cys Leu Glu Pro Pro Tyr Thr Gly Pro Cys Lys Ala Arg Ile Ile Arg Tyr Ala Tyr Asn Ala Lys Ala Gly Leu Cys Gln Thr Phe Val Tyr Gly Gly Cys Arg Ala Lys Arg Asn Asn Phe Lys Ser Ala Gleu Asp Cys Met Arg Thr Cys Gly Gly Ala

Or, using the IUPAC single-letter code:

RPDFCLEPPYTGPCKARIIRYAYNAKAGLCQTFVYGGCRAKRNNFKSAEDC MRTCGGA

Figure 1-1 shows a generalized amino acid residue, with the spheres representing individual atoms, and the letters within the circles identifying the type of atom. The circle marked "R" symbolizes the "side chain" which gives each of the twenty amino acids its individual identity. For now it is sufficient to note that each amino acid consists of a nitrogen atom bonded to an "alpha carbon" (the C atom in the center of the figure) that is in turn bonded to another carbon atom called the "carbonyl carbon". Proteins are created by a complex chemical or enzymatic process that fuses the amino acid monomers by creating what's called a "peptide bond" between the carbonyl carbon of one and the backbone nitrogen of the next. Thus, the N-C-C sequences of the individual amino acids are linked to form a continuous chain (illustrated in Figure 1-2) that constitutes the protein's "backbone". This linear assembly of residues is also known as the protein's primary structure.

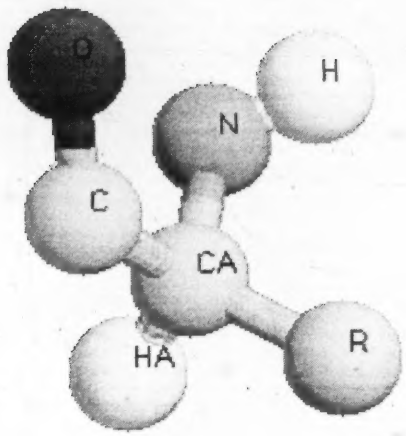


Figure 1-1: Generalized Amino Acid

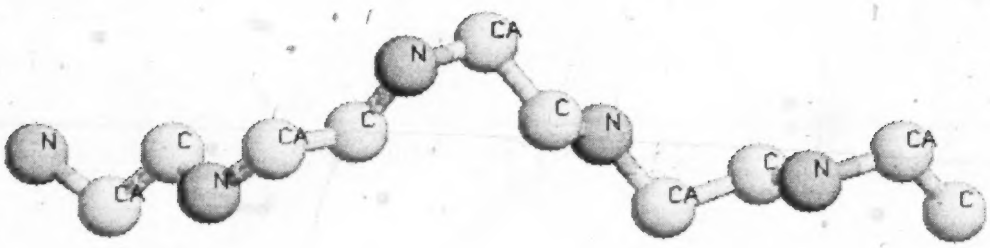


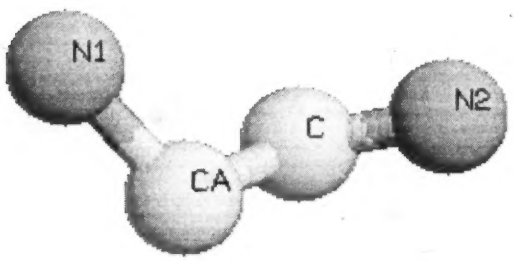
Figure 1-2: Protein Backbone

Once the protein has been synthesized from its constituent amino acids (using the translational machinery of the cell called the ribosome), it spontaneously "folds" into a specific three-dimensional shape – also called its tertiary structure Generally speaking, each unique amino acid sequence has a unique conformation into which it folds, and will always fold into that conformation (Lesk, 2001; Anfinsen, 1973). That conformation determines the protein's electrical and chemical characteristics, which in turn determine the protein's function.

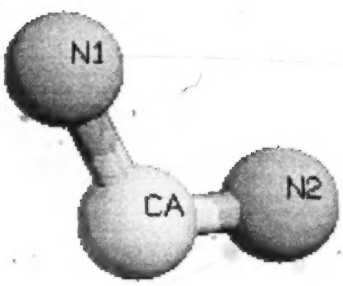
Protein Structures

How are these three-dimensional protein structures described and documented? Again, a list of atoms and their locations in space along with a list of bonds would suffice, but such a cartesian coordinate presentation fails in terms of emphasizing the "interesting" features of proteins over predictable information. As mentioned above, the sequence of atoms along the protein backbone does not vary between protein types; only the number of repetitions of the N-Ca-C sequence changes. Also, the atomic forces (van der Waals forces, double bond character of the peptide bond, coulomb forces in the sigma bonds as well as excluded volume effects) within the backbone only permit a limited range of variation in backbone geometry. In particular, most of the variation in backbone geometry takes the form of rotation around the two bonds connecting the nitrogen and the carbonyl carbon to the alpha carbon. To visualize this, imagine a section of the backbone four atoms in length (Figure 1-3a). Further imagine that you are looking at this section along the bond which joins the alpha and carbonyl carbons such that the alpha carbon is in front of you and hides the carbonyl carbon behind it (Figure 1-3b). Now what you see is the two nitrogen atoms and the bonds that join them to their respective carbons. The angle formed by these two N-C bonds from this point of view is a called a "torsion" or "dihedral" angle, and by convention is measured from the bond nearest the viewer to the bond furthest from the viewer, with clockwise rotation considered positive and counterclockwise as negative. In Figure 1-3b, the angle shown is about +115°. If the nitrogen atoms were to be exchanged, it would be -115°. The torsion angle measured looking along the alpha carbon-carbonyl carbon bond is called the "psi" angle. The angle

looking from the alpha carbon to the nitrogen is the "phi" angle. Simply listing the primary sequence, phi angles, and psi angles of a protein concisely conveys a great deal of information about that protein's structure.



A



B

Figure 1-3: Torsion Angles. a) The protein backbone viewed obliquely. b) The protein backbone viewed from a point in line with the C-C bond.

This kind of compact representation is called the "internal" coordinate representation and it is frequently displayed or depicted graphically in the form of a Ramachandran plot (Ramachandran and Sasisekharan, 1968), an example of which is shown in Figure 1-4.

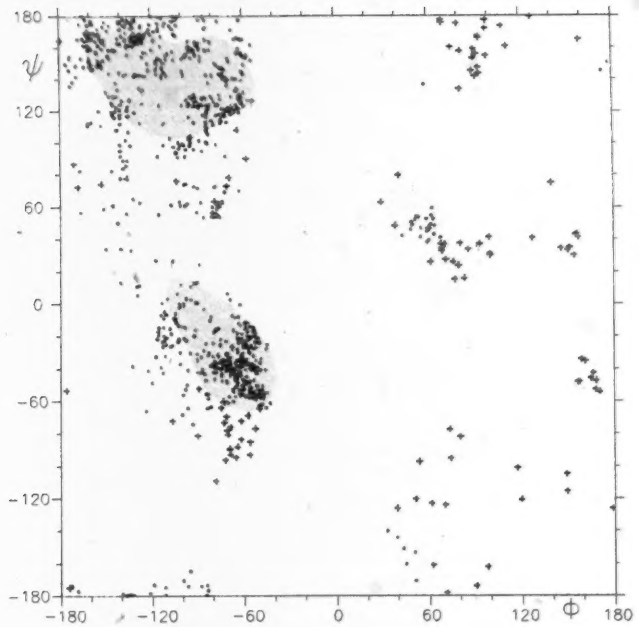


Figure 1-4: Ramachandran plot. Each residue is represented by a mark placed according to the phi and psi angles of the residue. The shaded areas represent regions where most phi/psi pairs should fall. Figure prepared with the program MOLMOL (Koradi, et al., 1996).

Also, as mentioned previously, each of the twenty amino acids has a unique side chain with its own physical characteristics. The varying properties of the side chains (length, geometry, atomic composition, flexibility, charge, hydrophobicity, etc.) are what allow proteins to achieve their diversity of structure and function. As with the backbone, the side chains are also defined by torsion angles (known by the Greek letter "chi"). Likewise, the chi angles can help describe amino acid side chain geometry in a concise manner (i.e. internal coordinate representation).

Hydrogen Bonds

Unlike the covalent peptide bonds which are responsible for maintaining the chemical integrity (and primary structure) of proteins, hydrogen bonds are actually weak non-

covalent bonds. They are primarily responsible for maintaining the tertiary structure or conformational integrity of proteins (Rose and Wolfenden, 1993). Hydrogen bonds form wherever a polar or partially polarized heavy atom (typically an oxygen atom) exerts a strong pull on a nearby hydrogen atom to which it is not covalently bonded. Typically the hydrogen atom actually shares its electron with the electronegative partner – as long as the distance between the two is less than about 2.5 Angstroms (Pimentel and McLellan, 1971; Ippolito et al., 1990). This "hydrogen bonding" helps stabilize proteins, and creates regions of structural regularity that further simplify the task of describing protein structure. Beyond the amino acid level or so-called primary structure level, proteins can be viewed as segmented collections of "secondary structure". These secondary structures fall into three classes:

- helical, in which the protein backbone takes a "corkscrew" shape;
- sheet, in which the protein backbone takes on an extended or ribbon-like shape which is paired (through hydrogen bonds) with one or more similarly extended strands to form a stable "sheets"; and
- coil, all other arrangements.

The most important hydrogen bonds in proteins form between the backbone amide hydrogen (which is bonded to the backbone nitrogen) and nearby backbone (or carbonyl) oxygen atoms. These kinds of backbone hydrogen bonds are important in forming the sheet and helix secondary structures (Lesk, 2001). Hydrogen bonds involving "alpha hydrogens" (hydrogen atoms attached to the alpha carbon) also form, but are much less

abundant than the amide hydrogen variety (Wagner et al. 1983; Wishart et al. 1991; Derewenda et al.,1995).

Disulfide Bonds

Another covalent bond that is important in protein structure (particularly in defining tertiary structure) is the disulfide or "two-cysteine" bond. Like peptide bonds, disulfide bonds are chemically stable covalent bonds. However unlike peptide bonds, disulfide bonds exist between side chains and not along the length of the backbone. Disulfide bonds can form spontaneously between any two cysteine residues that are close to one another in space (but potentially distant from one another in sequence). Disulfide bonds are quite strong once formed, and contribute strength and stability to a protein's structure, particularly secreted proteins that are exposed to environmental oxygen or other "rough" conditions. Note that disulfide bonds can be broken under appropriate reducing conditions and in fact, the interior or cytoplasm of most cells is sufficiently reducing to keep most disulfides from forming (Creighton, 1993).

The Significance of Protein Structure

The preceding description conveys important information about our current understanding of and methods for describing protein structure, but says little about the motivation for doing so. In brief, the primary reason for studying protein structure is that structure is the determinant of function. In other words, the shape, electrical, and chemical characteristics of a protein give rise to its function. As had been alluded to earlier, the importance of protein function is self-evident, especially when one considers

that many human disease states arise from disruptions of protein structure and function. Thus, to understand function at an atomic level, we are motivated to study protein structure at an atomic level. However, given that most proteins are too small (typically less than 10 nm in diameter) to be visualized directly (even using electron microscopes) special methods have had to be developed to examine their structures indirectly.

Protein Structure Determination

The oldest and most precise method of determining protein structures is X-ray crystallography (Lesk, 2001; Heinemann, Illing and Oschkinat, 2001; Wyckoff, Hirs, and Timasheff, 1985). In X-ray crystallography, an X-ray beam of a particular wavelength (or set of wavelengths) is passed through a solid crystal containing trillions of identically oriented protein molecules. The X-rays interact with and are diffracted by the electrons in the protein, yielding a complex diffraction pattern (typically a collection of thousands of spots) that may be detected by photographic film or an electronic area detector (similar in concept to a digital camera). The diffraction pattern or diffraction "shadow" thus cast can be analyzed mathematically (using Fourier analysis) to yield a three-dimensional contour map corresponding to the protein's electron density. Depending on the resolution of the resulting image, this electron density map can then be re-interpreted through careful visual analysis into the positions and coordinates for each of the atoms that comprise the protein – yielding a set of cartesian coordinates that fully defines the protein structure. A more complete description of the principles of X-ray crystallography can be found in Drenth (1994). The meaning of the term "resolution" as applied to protein structures will be explored in more detail below. For now, it is sufficient to say that

resolution is related to how detailed and accurate this electron density map is. The maximum *theoretical* resolution derivable by X-ray crystallography is determined by the protein crystal under study and the wavelength of the impinging X-rays (Creighton, 1993, p. 210). In *practice*, computational complexity may prove to be the limiting factor, as each increase in resolution requires a significant increase in mathematical effort.

More recently, NMR spectroscopy has emerged as an alternate (and complementary) approach to macromolecular (especially protein) structure determination (Wuthrich, 1986; Heinemann, Illing, and Oschkinat, 2001). In NMR spectroscopy the sample is typically kept in a liquid state (rather than a crystalline state) and placed in a powerful magnetic field (supplied by a superconducting magnet). Radio-frequency radiation is pulsed into the sample and the absorption bands (resonances) corresponding to individual atoms in the protein are recorded – yielding an NMR spectrum. Using specially designed combinations of radio frequency pulses with certain length, strength, and orientation, it is possible to simultaneously record the NMR spectra of multiple types of atoms (hydrogen, carbon, nitrogen), coupled atoms or different NMR parameters (NOEs, coupling constants, dipolar couplings, chemical shift anisotropy, etc.). Using two and three-dimensional NMR experiments such as TOCSY (Bax, 1989), NOESY (Kumar, Ernst, and Wuthrich, 1980), HSQC (Bodenhausen and Ruben, 1980) and HNCACB (Wittekind and Muller, 1993) it is possible to determine the chemical shifts, coupling constants, and spatial proximity of most NMR-active atoms in a protein. Once the chemical shifts have been assigned and the spatial proximity of most hydrogen atoms has been determined (through NOE – nuclear Overhauser effect – measurements), then it is possible to use computational techniques such as distance geometry (Wuthrich, 1986) or

simulated annealing (Brunger, 1993) to determine the 3D structure of the protein. As a rule, NMR spectroscopy is not as accurate as X-ray crystallography in terms of resolution or precision. However, NMR methods are quite effective for less structured, smaller proteins that are do not crystallize easily.

X-ray crystallography has been used since the late 1950's for protein structure determination, while NMR spectroscopy has been used since the early 1980's. Together these techniques have been used to determine the structure of more than 20,000 proteins and peptides, with X-ray crystallography accounting for about 16,000 structures and NMR accounting for the rest. Both NMR and X-ray crystallography are time-consuming and expensive. Typically it takes 1-2 years to solve a protein structure by NMR or X-ray crystallography (prior to ~1990, the average time was 5 years and prior to 1970 the average length of time was 10 years). The most time-consuming and "risky" aspect of X-ray analysis is the requirement for the creation of high quality protein crystals – which, for some proteins, is simply not possible. NMR is generally applied to proteins in solution and does not require crystals. However, NMR is less precise, requires expensive isotopic labeling and has limits on how large a protein it can analyze (typically less than 30 kilodaltons).

The limitations on size, precision and time to generate or assign a protein via NMR are still major bottlenecks. Many of these limitations arise out of the need for NMR spectroscopists to collect and assign extensive sets of NOE data and their almost complete reliance on NOE data (to the exclusion of many other structurally useful parameters) to provide necessary structural constraints for structure generation. These

NOE spectra are notoriously difficult to collect or interpret and are almost impossible to use when the size of the protein exceeds 25 kilodaltons (about 200 residues).

It is our contention that if chemical shifts could be more fully exploited in NMR that many of these bottlenecks relating to structural precision, size and time to assign or generate structures could be reduced or even eliminated. This is the primary reason for developing the protein chemical shift calculator described in Chapter 3 of this thesis.

In order to understand why there should be any correlation between protein structure and chemical shifts, a deep understanding of the NMR phenomenon is not required, but some basics will be useful in understanding the balance of the present work.

Chemical Equivalence

One concept that is very useful for understanding NMR is "chemical equivalence". Two atoms are said to be chemically equivalent if there is no way that they can be distinguished by examining the molecule or molecules to which they are attached. The two hydrogen atoms bonded to an oxygen atom in a water molecule are chemically equivalent; the symmetry of the molecule guarantees that there is no way to distinguish them. Likewise, all the hydrogen atoms on all the water molecules in a glass of water are chemically equivalent. More to the point, in a sample of a billion protein molecules, all the corresponding atoms on each molecule are chemically equivalent. This is important for understanding NMR because all chemically equivalent atoms respond in the same way to NMR stimulus.

Chemical Shifts

As stated earlier, when a sample is immersed in a powerful magnetic field and subject to a pulse of radio frequency electromagnetic (EM) energy it will absorb certain frequencies of the impinging EM radiation. Each NMR-detectable nuclei (hydrogen, carbon, and nitrogen being of particular interest in protein NMR) has a characteristic EM frequency ω at which it absorbs, which is dependent on a characteristic of the nucleus called the gyromagnetic ratio γ, and the strength of the magnetic field **B** in which the nucleus is immersed:

1-1) $\omega = \gamma B$

Isolated atoms in vacuum would be subject only to the primary magnetic field generated by the NMR apparatus, but the fields "experienced" by atoms in molecules are modified by the atom's local electromagnetic environment. This environment may "shield" or "deshield" the atom, respectively diminishing or increasing the strength of the effective magnetic field, thereby changing the atom's resonant frequency. A shielding constant σ exists for each atom, modifying the magnetic field:

1-2) $B_{eff} = B_o(1-\sigma)$

where B_o is the applied magnetic field, and B_{eff} is the effective magnetic field experienced by the nucleus. A number of factors determine the shielding or deshielding of an atom, most notably the electron cloud surrounding the nucleus, which in turn is strongly influenced by the atom's covalent bonds and bonding partners. Electrostatic charges, ring currents, solvents, paramagnetic moieties, and nearby (non-bonded) nuclei also have substantial effects on shielding. These factors combine in a non-additive, non-

linear manner (Wuthrich, 1986; Clore and Gronenborn, 1993). An example of this complexity is given in Figure 1-5. The environments of the atoms shown in 1-5a and 1-5b differ only slightly (by the placement of one chemical bond), but their chemical shifts may be dramatically different. The atom shown in Figure1-5c has an environment is the combination of those in 1-5a and 1-5b, but its chemical shift will likely be very different from the sum of the shifts of those atoms.

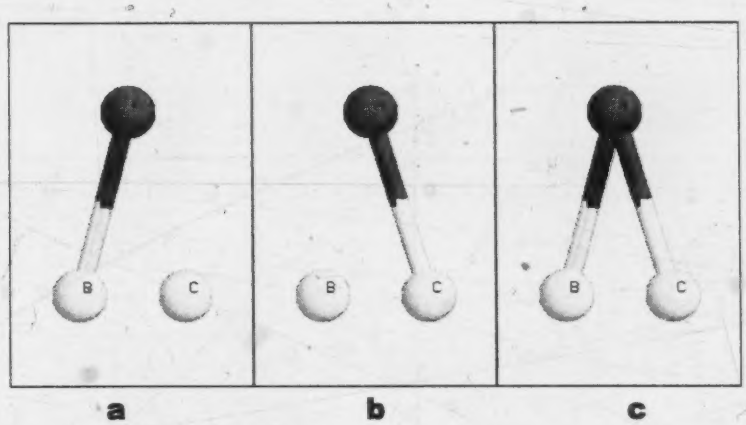


Figure 1-5: Atomic Bonding Environments.

Different nuclei (or atom types) will respond to different electronic influences or chemical-bonding environments in different ways. Some atoms or nuclei (namely hydrogen atoms) are highly sensitive to hydrogen bonds and the presence of aromatic rings. Other atoms or nuclei (namely carbons) are more sensitive to covalently bonded substituents or bond angles. Solvent effects, salt effects, paramagnetic ions, and other hard-to-classify "environmental" phenomena can affect all nuclei but the magnitude of their influence varies (Wishart and Case, 2001; Clore and Gronenborn, 1994).

A general equation for shielding may be written as:

$$\text{1-3)} \qquad \sigma = \sigma_{dia}(local) + \sigma_{para}(local) + \sigma_{anis} + \sigma_{rc} + \sigma_{ef} + \sigma_{solv}$$

enumerating diamagnetic, paramagnetic, anistropic, ring current, electric field, and solvent effects respectively. Each of these factors will be discussed in more detail below. In the more general case, one must also consider the effect of paramagnetic moieties (those with unpaired electrons) in the molecular system under study, but this project specifically excluded such systems.

Diamagnetic Effects

Atomic nuclei are surrounded by clouds of electrons, which respond to an applied magnetic field by precessing about the axis of the magnetic field. This creates an electric current and associated magnetic field that opposes the applied magnetic field. The shielding effect σ can be in theory be calculated:

$$\text{1-4)} \qquad \sigma = \frac{\mu_o e^2}{3m_e} \int_0^\infty r\,\overrightarrow{\rho(r)}dr$$

where m_e is the mass of an electron, r is the distance from the electron to the nucleus, ρ(r)is the electron density, e is the electrical charge of an electron, and μ_o is a constant of proportionality. However, this formula applies only to the special case of electron spherical symmetry; the general case is far more complex.

Paramagnetic Shielding (Local)

The paramagnetic term accounts for departure from spherical electron rotation, caused by the formation of bonds among other things. Calculating the shielding from this

effect requires knowledge of how the electron wave functions are impacted by the local magnetic field, and is beyond the scope of this introduction.

Anisostropic Shielding

This term accounts for chemical shift changes due to different molecular orientations relative to the applied magnetic field. If the molecules of a sample are free to rotate, this term is not significant, but it becomes important in solid-state NMR and time-domain NMR observations.

Ring Currents

Molecular rings incorporating electrons in pi bonds create electron currents around the ring that give rise to magnetic fields oriented along the ring's axis. The effects of this field on chemical shifts are quite significant, and may be shielding or deshielding, depending on the position of the nucleus under study relative to the ring axis. The amino acids tyrosine, tryptophan, and phenylalanine contain such rings, and need to be taken into account when calculating the chemical shifts of nearby atoms. Many different ways of mathematically modeling this effect have been proposed (Haigh and Mallion, 1980). Chapter 3 of this thesis details the calculation of ring current effects.

Electrostatic Effects

Charged groups affect chemical shifts most directly by polarizing nearby bonds and otherwise distorting electron density in the molecule. Also, bonds between nuclei of

different electronegativity will also be polarized, deshielding the atom from which electrons are withdrawn.

Solvent Effects

Based on the foregoing, it is easy to see that the properties of the solvent can have a profound effect on chemical shifts. A polar solvent (like water) or one incorporating rings (like benzene) will profoundly affect the shifts of solvent-accessible atoms. Changing solvents can thus be a useful tool for analyzing protein structure.

Theory and Mathematical Prediction of Chemical Shifts

It is clear that the relationship between atomic environment and chemical shifts is very complex, and as mentioned previously, is dependent on many factors in a non-linear manner. It is, however, possible to calculate chemical shifts very precisely from structural data, using the techniques of both classical and quantum mechanics. Unfortunately, the computation required to solve even a very simple atomic system is extensive. Indeed, for all but the simplest of real-world systems, the required quantum mechanical calculations are computationally intractable, even with modern supercomputers. The present work is concerned with predicting chemical shifts computationally, but using a combination of equations derived from classical physics, and statistically-derived tables (also called "hypersurfaces").

Random Coil Shifts

Under certain chemical conditions, proteins lose their native 3D structure and fall into a disordered state called "random coil", in which the entire chain of amino acids is free to "wiggle" as it is buffeted by atomic collisions. The chemical shift of an atom in such a random coil protein is effectively the chemical shift averaged over a random assortment of atomic environments, with only the nearest rigidly-bonded atoms remaining in constant positions relative to the atom of interest. Random coil chemical shifts are therefore used as the starting point in chemical shift predictions, being effectively the chemical shift of the atom without the influence of protein's native structure. Random coil shifts have been determined experimentally by a number of authors (see, for example, Bundi and Wuthrich, 1979; Keim et al., 1973; Merutka et al., 1995; Wishart et al., 1995). Typically these involve the measurement of chemical shifts of defined penta or hexapeptides containing the residue of interest in the middle of the sequence. Note that residues random coil shifts are substantially different than single amino acid shifts and that single amino acid shifts are not particularly useful for interpreting the chemical shifts of polypeptides (which are composed of covalently linked amino acid residues). For the purposes of this thesis and the work described in Chapters 2 and 3 we have used the amino acid residue random coil shifts provided by Wishart et al. (1995). These are shown in Table 1-2.

Amino Acid	CO	CA	CB	HN	HA	HB	Side chain H
Ala	177.8	52.5	19.1	8.24	4.32	1.39	
Cys (reduced)	174.6	58.2	28.0	8.32	4.55	2.93, 2.93	
Cys (oxidized)	174.6	55.4	41.1	8.43	4.71	3.25, 2.99	
Asp	176.3	54.2	41.1	8.34	4.64	2.72, 2.65	
Glu	176.6	56.6	29.9	8.42	4.35	2.06, 1.96	γ: 2.31, 2.31
Phe	175.8	57.7	39.6	8.30	4.62	3.14, 3.04	2,6H: 7.28 3,5H: 7.38 4H: 7.32
Gly	174.9	45.1		8.33	3.96		
His	174.1	55.0	29.0	8.42	4.73	3.29, 3.16	2H: 8.58 4H: 7.29
Ile	176.4	61.1	38.8	8.00	4.17	1.87	γCH_2: 1.45, 1.16 γCH_3: 0.91 δCH_3: 0.86
Lys	176.6	56.2	33.1	8.29	4.32	1.84, 1.75	γCH_2: 1.44, 1.44 δCH_2:1.68, 1.68 εCH_2:2.99, 2.99 εNH_3^+ 7.81
Leu	177.6	55.1	42.4	8.16	4.34	1.62, 1.62	γCH 1.59 δCH_3:0.92, 0.87
Met	176.3	55.4	32.9	8.28	4.48	2.11, 2.01	γCH_2 2.60, 2.54 εCH_3: 2.10
Asn	175.2	53.1	38.9	8.40	4.74	2.83, 2.75	γNH_2 7.59, 6.91
Pro	177.3	63.3	32.1		4.42	2.29, 1.94	γCH_2 2.02, 2.02 δCH_2:3.63, 3.63
Gln	176.0	55.7	29.4	8.32	4.34	2.12, 1.99	γCH_2 2.36, 2.36 δNH_2:7.52, 6.85
Arg	176.3	56.0	30.9	8.23	4.34	1.86, 1.76	γCH_2:1.63, 1.63 δCH_2:3.20, 3.20 εNH: 8.07
Ser	174.6	58.3	63.8	8.31	4.47	3.89, 3.87	
Thr	174.7	61.8	69.8	8.15	4.35	4.24	γCH_3: 1.21
Val	176.3	62.2	32.9	8.03	4.12	2.08	γCH_3:0.94, 0.93
Trp	176.1	57.5	29.6	8.25	4.66	3.29, 3.27	2H: 7.27 4H: 7.65 5H: 7.18 6H: 7.25 7H: 7.50
Tyr	175.9	57.9	38.8	8.12	4.55	3.03, 2.98	2,6H: 7.14 3,5H: 6.84

Table 1-2: Random coil shifts of amino acids. These are the chemical shifts that would be expected for residues in a disordered polypeptide.

Databases

Of the many databases of biomolecular information available, two were of particularly importance in this study. The most important source of protein structural information is the Protein Data Bank, generally referred to as the PDB (Berman et al., 2000), which can be accessed at the URL http://www.pdb.org. At time of this writing, the PDB contained 19688 protein structures. Of these, 2687 were determined using NMR, and the remainder with X-ray crystallography. This organization, in addition to being the curators of the world's largest database of protein structures, has also defined a standard computer file format for recording and exchanging protein structures; this format is naturally known as the "PDB format". Some example records from a PDB file are shown in Figure 1-6.

```
HEADER    PROTEINASE INHIBITOR (TRYPSIN)        30-APR-92   1PIT
COMPND    TRYPSIN INHIBITOR
SOURCE    BOVINE (BOS TAURUS) PANCREAS
EXPDTA    NMR
AUTHOR    K.D.BERNDT,P.GUNTERT,L.P.M.ORBONS,K.WUTHRICH
JRNL      AUTH   K.D.BERNDT,P.GUNTERT,L.P.M.ORBONS,K.WUTHRICH
JRNL      TITL   DETERMINATION OF A HIGH-QUALITY NUCLEAR MAGNETIC
JRNL      TITL 2 RESONANCE SOLUTION STRUCTURE OF THE BOVINE
JRNL      TITL 3 PANCREATIC TRYPSIN INHIBITOR AND COMPARISON WITH
JRNL      TITL 4 THREE CRYSTAL STRUCTURES
JRNL      REF    J.MOL.BIOL.                V. 227   757 1992
JRNL      REFN   ASTM JMOBAK  UK ISSN 0022-2836          070
SEQRES  1    58  ARG PRO ASP PHE CYS LEU GLU PRO PRO TYR THR GLY PRO
SEQRES  2    58  CYS LYS ALA ARG ILE ILE ARG TYR PHE TYR ASN ALA LYS
SEQRES  3    58  ALA GLY LEU CYS GLN THR PHE VAL TYR GLY GLY CYS ARG
SEQRES  4    58  ALA LYS ARG ASN ASN PHE LYS SER ALA GLU ASP CYS MET
SEQRES  5    58  ARG THR CYS GLY GLY ALA
HELIX   1  H1 ASP     3  GLU     7  5 ALL DONORS,ACCEPTORS INCLUDED
HELIX   2  H2 SER    47  GLY    56  1 ALL DONORS,ACCEPTORS INCLUDED
SHEET   1  S1 3 LEU   29  TYR    35  0
SHEET   2  S1 3 ILE   18  ASN    24 -1  N  ILE    18   O  TYR    35
SSBOND  1 CYS    5    CYS    55
SSBOND  2 CYS   14    CYS    38
ATOM     1  N   ARG     1      -8.544   3.578  14.046  0.00  2.37
ATOM     2  CA  ARG     1      -7.776   3.484  12.790  0.00  1.79
ATOM     3  C   ARG     1      -8.492   4.333  11.742  0.00  1.59
ATOM     4  O   ARG     1      -9.713   4.432  11.844  0.00  1.71
ATOM     5  CB  ARG     1      -7.640   2.025  12.309  0.00  1.57
```

Figure 1-6: Records from a PDB file.

A PDB file consists of a set of records (one record per line), each of which has a well-defined format. In Figure 1-6, the first several records are bibliographic in nature, describing the substance under study, the investigators, the journal of publication, and the experimental details. Other records describe the physical characteristics of the protein: SEQRES records list the amino acid sequence, HELIX and SHEET records describe regions of secondary structure, and SSBOND records indicate the presence of disulfide bonds. Most important are the ATOM records, which give the coordinates of each atom (in units of Angstroms) relative to an arbitrary center. A full description of the file format may be found at the URL:

http://www.rcsb.org/pdb/info.html#File_Formats_and_Standards

Atomic Nomenclature in Proteins

The PDB has also developed some conventions on atomic nomenclature to ensure consistency. Each atom in each amino acid has a standard name associated with it, identifying the type of atom, and its position within the amino acid. Hydrogen names all begin with "H", carbons with "C", nitrogens with "N", oxygens with "O", and sulfur atoms with "S". The backbone atom names are the same for all amino acids: N for the backbone amide nitrogen, H for the amide hydrogen, CA for the alpha carbon, HA for the alpha hydrogen, C for the carbonyl carbon, and O for the carbonyl oxygen. As mentioned elsewhere, the alpha carbon is the central carbon in the amino acid backbone; the side chain atoms are bonded to it and labeled with the consecutive symbols of the Greek alphabet: beta, gamma, delta, epsilon, zeta, and eta, or, as they are abbreviated: B, G, D,

E, Z, H. Table 1-3 lists the sidechain protons of each of the amino acids; more details can be found on the BioMagResBank website, the location of which can be found below.

Chemical Name	Sidechain Protons
Alanine	HB1, HB2, HB3
Cysteine	HB2, HB3, HG
Aspartic Acid	HB2, HB3, HD2
Glutamic Acid	HB2, HB3, HG2, HG3, HE2
Phenylalanine	HB2, HB3, HD1, HD2, HE1, HE2, HZ
Glycine	HA2, HA3
Histidine	HB2, HB3, HD1, HD2, HE1, HE2
Isoleucine	HB, HG21, HG22, HG23, HG12, HG13, HD11, HD12, HD13
Lysine	HB2, HB3, HG2, HG3, HD2, HD3, HE2, HE3, HZ1, HZ2, HZ3
Leucine	HB2, HB3, HD11, HD12, HD13, HD21, HD22, HD23, HG
Methionine	HB2, HB3, HG2, HG3, HE1, HE2, HE3
Asparagine	HB2, HB3, HD21, HD22
Proline	HB2, HB3, HD2, HD3, HG2, HG3,
Glutamine	HB2, HB3, HG2, HG3, HE21, HE22
Arginine	HB2, HB3, HG2, HG3, HD2, HD3, HE, HH11, HH12, HH21, HH22
Serine	HB2, HB3, HG
Threonine	HB, HG1, HG21, HG22, HG23
Valine	HB, HG11, HG12, HG13, HG21, HG22, HG23
Tryptophan	HB2, HB3, HD1, HE1, HE3, HZ2, HZ3, HH2
Tyrosine	HB2, HB3, HD1, HD2, HE1, HE2, HH

Table 1-3: Sidechain protons

Structure Resolution

PDB files for X-ray structures also usually specify a "resolution" which is related to the accuracy with which atomic positions are reported; resolutions are typically given in units of Angstroms, an Angstrom (abbreviated Å) being 10^{-10} meters. To say that a protein structure has been determined to 2 Angstroms resolution suggests (the implication is not exact) that the one can look at the structure and distinguish the positions of any two atoms that are two or more Angstroms apart. Atoms that are closer together may appear to be one large feature. The study detailed elsewhere in this document used only high quality X-ray with resolutions of 2 Å or greater.

The BioMagResBank

The other key database for the present study is the BioMagResBank, often abbreviated as BMRB (Seavey et al., 1991), located at the URL http://www.bmrb.wisc.edu. As the name implies, it is a databank of biomolecular NMR data, and contains the measured chemical shifts for many different proteins (currently 559392 shifts). This is the key way that protein NMR researchers share their results.

Outline and Objectives of this Dissertation

Simply stated, there are two objectives to this thesis: 1) armed with a set of high resolution protein structures and the chemical shifts of atoms in those structures, use data mining techniques to determine what aspects of those protein structures are important in determining the chemical shifts; and 2) with that determination made, write a computer program that will accept as input a PDB-format protein structure, and output a list of accurately predicted chemical shifts (^{1}H, ^{13}C and ^{15}N). The first objective is the goal of MINER, a computer program to facilitate data mining, which is documented in Chapter 2. The second objective (to accurately and rapidly calculate ^{1}H, ^{13}C and ^{15}N chemical shifts in diamagnetic proteins) is the province of a second computer program called SHIFTX, which is detailed in Chapter 3. Chapter 4 of this thesis summarizes some of the key results from Chapters 2 and 3 as well as outlining some future research goals that could be undertaken in light of these results.

References

Anfinsen, C.B. (1973) *Science*, **181**, 223-230

Bax, A.(1989) *Meth. Enzymol.*, **176**, 151-168

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Research*, **28,** 235-242.

Bodenhausen G. and Ruben, D. J. (1980) *Chem. Phys. Lett.*, **69**, 185-189

Brunger, A.T. (1993) *X-Plor Version 3.1: A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, CT

Bundi, A., and Wuthrich, K. (1979) *Biopolymers*, **18**, 285-297.

Case, D.A. (2000) *Curr. Opin. Struct. Biol.* **10**, 197-203.

Clore, G.M. and Gronenborn, A.M. eds. (1993) *NMR of Proteins*, CRC Press, Boca Raton, FL

Creighton, T. (1993) *Proteins. 2nd ed.* W.H. Freeman and Company, New York, NY

Dean, P.M., Zanders, E.D., and Bailey, D.S. (2001) *Trends Biotechnol.*, **19**, 288-292

Derewenda, Z.S., Lee, L. and Derewenda, U. (1995) *J. Mol. Biol.* **252**, 248-262.

Drenth, J. (1994) *Principles of Protein X-ray Crystallography*. Springer-Verlag, New York, NY and London.

Haigh, C.W. and Mallion, R.B. (1980) *Progress In NMR Spectroscopy*, **13**, 303-344.

Heinemann, U., Illing, G. and Oschkinat, H. (2001) *Curr. Opin. Biotechnol.*, **12**, 348-354

Ippolito, J.A., Alexander, R.S. and Christianson, D.W. (1990) *J. Mol. Biol.*, **215**, 457-471

Keim, P., Vigna, R.A., Marshall, R.C. and Gurd, F.R.N. (1973) *J. Biol. Chem.*, **248**, 7811-7818

Kumar, A., Ernst, R.R., and Wuthrich, K. (1980) *Biochem. Biophys. Res. Com.* **95**, 1-6.

Koradi, R., Billeter, M., and Wüthrich, K. (1996) *J. Mol. Graphics* **14**, 51-55.

Lesk, A. (2001) *Introduction to Protein Architecture*, Oxford University Press, New York, NY.

Maggio, E.T. and Ramnarayan, K. (2001) *Trends Biotechnol.*, **19**, 266-272

Merutka, G., Dyson, H.J. and Wright, P.E. (1995) *J. Biomol. NMR*, **5**, 14-24

Mitchell, T.M. (1997) *Machine Learning*, McGraw-Hill, New York, NY.

Oldfield, E. (1995) *J. Biomol. NMR*, 5, 217-25.

Pimentel, G.C. and McLellan, A.L. (1971) *Rev. Phys. Chem.*, **22**, 347-385

Ramachandran, G.N. and Sasisekharan, V. (1968) *Adv. Protein Chem.*, **23**, 283-438

Ramakrishnan, N. and Grama, A.Y. (1999) *Computer*, **32**, 34-37.

Rose, G.D. and Wolfenden, R. (1993) *Annu. Rev. Biophys. Biomol. Struct.*, 22, 381-415

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217-236.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490-5492.

Szilagyi, L. (1995) *Prog. Nucl. Magn. Reson. Spect.* **27**, 325-443

Wagner, G., Pardi, A. and Wuthrich, K. (1983) *J. Am, Chem. Soc.*, **105**, 5948-49.

Williamson, M.P. and Asakura, T. (1997) *Meth. Mol. Biol.*, **60**, 53-69

Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR*, **6**, 135-40.

Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3-34

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol. Biol.*, **222**, 311-333.

Wittekind, M. and Muller, L. (1993) *J. Magn. Reson. B* **101**, 201-205

Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York, NY

Wyckoff, H.W., Hirs, C.H.W. and Timasheff, S.N. eds. (1985) *Meth. Enzymol.*, **114-115**

Xu, X-P. and Case, D.A. (2002) *Biopolymers*, **65**, 408-423.

Chapter 2: MINER

Introduction

Chapter 1 summarized some of the experimentally observable factors that can be used to predict chemical shifts. These include structural parameters (phi, psi, and chi), the existence of bonds and their lengths, secondary structure, and several other factors. One of the key tasks in developing algorithms for chemical shift prediction was not just finding which factors were important for predicting chemical shifts, but finding which factors interact with other factors, such that these combined factors could be more effective at making predictions than the individual factors alone.

Consider a hypothetical "ideal" situation where one has a large quantity of error-free data and perfect knowledge of all factors that affect chemical shifts. It might be possible to create an analytical formula that would yield an exact chemical shift for any given set of input variables. However, even if one could not create a closed-form and computationally feasible formula to calculate chemical shifts, one could instead create a multi-dimensional table with an axis for each factor. With this table in hand, chemical shift prediction would become a simple matter of looking up the appropriate cells in the table.

Unfortunately, neither of these conditions prevails in biomolecular NMR. There are errors in the data, and the influence of the various factors is not well understood. The approach taken in the present work was to approximate the "ideal" multi-dimensional

table by a collection of one- and two-dimensional tables, accepting an increase in prediction error in exchange for ease of computation.

Earlier workers in the field have identified several factors and factors pairs that strongly influence chemical shifts (see references in Chapters 1 and 3) but no systematic search had been made to identify all factors and pairs statistically useful in predicting shifts. MINER, the subject of the current chapter, is a data mining computer program written specifically to make such a search. It is important to note that creating this program was much more of an engineering project than a scientific one. The procedures and statistics used are relatively simple and well understood; the challenge was combining them in a program that produced correct results in a reasonable timeframe. This engineering orientation is readily seen in this chapter and its emphasis on finding working solutions rather than understanding why they work.

Methods

MINER was written in the C programming language, compiled using Microsoft Visual C++ (versions 6 and 7) under all post-95 32-bit versions of the Windows operating system running on standard Pentium-type PCs, operating at clock speeds varying from 600 MHz to 2.4 GHz. A version was also compiled using GCC and tested on the Solaris operating system.

Two databases of protein structures and the corresponding chemical shifts were used as input to MINER. One database was for backbone ^{1}H and heteronuclear (^{13}C and ^{15}N) shifts, and the other was for ^{1}H side chain shifts. The database used to calibrate backbone ^{1}H, ^{13}C and ^{15}N shift predictions consisted of 37 diamagnetic proteins assembled from an extensive search of the literature and BioMagResBank (Seavey et

al.,1991). These proteins, their BMRB and PDB accession numbers as well as their resolution are shown in Table 3-3. In compiling this database the focus was on finding proteins that:

- were reliably referenced, as indicated from the literature or BMRB data;
- spanned a variety of structural classes (all α, all β, mixed α/β);
- were well-structured;
- had no "shift-significant" ligands or paramagnetic moieties; and
- had high resolution X-ray structures (<2.1 Å).

Data was selected from a variety of labs to prevent any instrumental or operator bias from creeping into the chemical shift calculations and the refinement process. Several sets of protein shifts (esp. ^{13}C and ^{15}N) were re-referenced to conform to IUPAC recommendations (Wishart et al., 1995; Markley et al., 1998). Glycine ^{1}HA shifts were averaged and treated as a single shift because of the limited information on stereospecific assignments.

The second database, used to analyze the side chain hydrogens, was developed by cross-referencing a list of BMRB shifts for nuclei of interest with a list of high resolution (2.0 Angstroms or greater) X-ray crystal structures from the Protein Data Bank (Berman et al., 2000). The selection criteria as described for the backbone database above was also employed for this database except that chemical shift referencing was not an issue for these shifts. Table 3-4 lists the BMRB and PDB files used for the side chain ^{1}H shift database.

Analytical Approach

It would naturally have been preferable to use well-understood and well-documented techniques to analyze the NMR/structural data, but no existing technique seemed suitable. There is a large body of literature concerned with interpolation, extrapolation, and approximation of many-dimensional functions, but most extant techniques require making one or more assumptions that do not hold in the present case:

- that the data is free of errors;
- that the function may be sampled at arbitrary points, or at least at regular intervals; or
- that a distance metric (a formula for calculating the distance between points) is available

This last is particularly troublesome, and bears some careful consideration. The most basic requirement of any interpolation technique is a means to decide which known points are closest to the desired point of interpolation. If points are described strictly in terms of continuous real numbers, formulating a distance metric requires a relatively straightforward application of geometry, possibly complicated by different weighting for different dimensions. The data used for MINER, however, includes not only real-valued attributes, but also discrete attributes, such as amino acid type. This raises a series of dilemmas along the lines of: given three residues with the same phi/psi angles, which is more similar to a cysteine: an alanine or a methionine? Of course, a table of distances could be computed from the data, but the way to do this without assuming a conclusion is

far from clear. Combinatorial risks also abound: a distance measure for amino acid type would require solving for 20x20=400 unknowns. If one was concerned that ignoring nearest neighbours would cause inaccuracies, one could instead compute distances for dipeptides (160,000 unknowns) or tripeptides (64,000,000 unknowns). It turns out that the information we are seeking (which factors are important in determining chemical shifts) is the same information we need to construct a distance metric. Lack of such a metric rules out the use of many otherwise promising techniques such a k-nearest-neighbours (Langley, 1996). It was with this in mind that the approach used in MINER was evolved: guided by the data, search for a linear combination of low-dimension estimators that forms a reasonable approximation to the unknown function.

Other techniques that have shown great promise for analyzing complex data, such as neural networks (Rumelhart and McClelland, 1986; Haykin, 1994) were not considered because they are "black boxes" that might yield good results but give no insight into how those results are computed. Since an important goal of this project was to gain insight into how various protein factors interact and influence chemical shifts, such techniques were not suitable.

From the outset, it was decided that only combinations of two factors would be studied; increasing the analysis beyond two dimensions would make the data sparser (and therefore less reliable) and vastly increase the computational burden. The key consideration then became finding which pairs of factors were most effective in predicting chemical shifts. Some pairings were obvious; the phi and psi angles together specify the relative positions of atoms in adjacent amino acids much more accurately than either angle alone. However, due to the statistical effects caused by collapsing a multi-

dimensional table into a one- or two-dimensional table, some of the factor combinations might be considerably less obvious.

A further issue needing consideration was how to evaluate which factor pairs were most useful for predicting chemical shifts. Given a collection of experimentally-measured chemical shifts and a corresponding set of predicted shifts, the prediction quality is reflected by one of two values. The first is the root-mean-squared deviation (RMSD), which is the average geometric distance between two sets of points. Given two sets of n points x and y, their RMSD is defined as (Press et al., 1992):

$$2\text{-}1) \qquad rmsd \equiv \sqrt{\frac{\sum_n (x_n - y_n)^2}{n}}$$

The other quality measure is the statistical value known as Pearson's Correlation Coefficient, or more simply, the "correlation". Given the same sets x and y, their correlation r is defined as (Press et al., 1992):

$$2\text{-}2) \qquad r \equiv \frac{\sum_n (x_n - \bar{x})(y_n - \bar{y})}{\sqrt{\sum_n (x_n - \bar{x})^2} \sqrt{\sum_n (y_n - \bar{y})^2}}$$

Both these values are easily and quickly computed, and provide an easy way to rank sets of predictions.

Using the ideas above, the program MINER was created to automatically evaluate each factor and pair of factors, select the most effective factor, apply its predictions to the data set, then repeat this process until no further improvement in prediction quality was

possible. In effect, MINER is a tool for the automatic construction and evaluation of tables (as defined below).

The factors used in the chemical shift predictions are listed in Table 3-2; their physical meanings were explained in Chapter 1 of this document. The input MINER consists of a large table, with each row containing an experimentally determined chemical shift and each column containing a value for one of the factors listed in Table 3-2. MINER then uses this data to construct a series of trial tables and evaluate their effectiveness.

Tables

This chapter will make frequent reference to "tables" and uses this term in an unusual enough way to call for an explicit definition. In most cases, the tables referred to herein consist of two-dimensional arrays of chemical shifts with the rows and columns arranged according to the value of one of the physical factors. An example is shown in Table 2-1, where each row is labeled with a particular type of secondary structure, and each column represents a range of psi values. The value found at a given row and column in the table will be referred to as a "cell"; for example, the top left cell in Table 2-1 contains the value "9.95", and is associated with psi values between –180 and –160 and a secondary structure type of "helix". Generally speaking, the present work deals only with two-dimensional tables, but note that the concept can readily be extended to three or more dimensions as required. The term "hypersurface" will frequently be used in this document to refer to two-dimensional tables.

		PSI				
		-180 -165	-165 -150	-150 -135	-135 -120	-120 -105
Secondary Structure	Helix	9.95	9.41	8.9	8.56	8.13
	Sheet	9.75	9.21	8.7	8.36	7.93
	Coil	9.55	9.01	8.5	8.16	7.73

Table 2-1: Example Table

The Best Guess

A more accurate statement of MINER's chief task is *to minimize the difference between predicted and actual chemical shifts.* This being the case, MINER starts each iteration through the data set with two chemical shift values for each of its input points: the actual shift, and a "best guess" value. The values that MINER places in tables (and attempts to minimize) are the differences between these two chemical shifts. The initial best guess consists of the sum of the physical factors (random coil, electrostatic effects, ring current effects, and disulfide bond factors, the computation of which values will be explained in the next chapter). The best guess on subsequent iterations will be the physical factors plus contributions from whichever tables MINER has discovered to be the most effective predictors (which is to say the most effective at minimizing errors).

Choosing Axes

Having generated a list of chemical shift error values in the previous step, MINER then places each of these in a cell in a table with axes using the selected factors. For example, consider a phi/amino acid table (Table 2-2). For discrete factors (for instance, amino acid type), the assignment of an axis is straightforward: one column or row for each of the twenty amino acid types. In some cases, these discrete axes would be binary,

indicating the presence or absence of a feature; the disulfide bond value would be represented by such an axis.

For a continuous value like phi, however, values must be chosen as the boundaries between one row and the next. In the example shown in Table 2-2, the phi region has been divided into three ranges of 120°. In fact, a more complex arrangement was empirically arrived at for angles in MINER, utilizing overlapping ranges 40° in width with their centers separated by 20°. This may seem an overly coarse division until one considers that uncertainty in atomic positions of as little 0.20Å implies uncertainty of more than 20° in torsion angles. Given the resolution of the structures used in this research, the 40° intervals adequately reflect the precision of the input data. Using this scheme, the first row in a hypothetical table would be labeled "–180°" and would contain data from points with phi values in the range –200° to –160° (or equivalently 160° to 200°). The next row would be labeled "–160°" and contain data corresponding to phi values from –160° to –180°, and so on until the last row, labeled "180°" (and thus identical to the "–180°" row).

Other continuous non-angle values (for example, bond lengths) are divided up into twenty equal-sized ranges (the value twenty having been reached empirically), with the lowest value present set to be the lower boundary of the leftmost column, and the highest value present in the data forming the upper boundary of the rightmost column. An additional complication arises for zero values in the data, which in the case of bond lengths represent the absence of a bond. For these situations, an additional "zero" column is added to the table, as opposed to creating a column containing points with bond lengths from zero to some intermediate value. This latter scheme would lump together bond-

absent with short-bond data points – leading to an undesirable mixing of atoms in quite distinct physical states.

		Amino Acid Type																		
		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Phi Angle	-180 to -60																				
	-60 to +60																				
	+60 to 180																				

Table 2-2: An example MINER table. "Amino Acid Type" is the horizontal axis and "Phi Angle" is the vertical.

Thus, the axes in MINER's tables consist of one of four types:

- discrete;
- continuous angle;
- continuous value, no zero bin; or
- continuous value with zero bin.

Building Tables

Having selected two factors for trial and constructed axes from the data, building the table is a simple matter of assigning each data point to a cell or cells in the table, and finding the mean of the error values assigned to each cell. In some runs of MINER, it was required that a cell contain a given number of data points (generally five) or it would be treated as empty. This requirement was to minimize the influence of outliers and erroneous data. Adding this restriction was not found to have any significant impact on

the results. Evidently, points in sparsely populated regions of the data space did not have widespread influence over the shape of the hypersurface.

Thus MINER generates a table containing all information necessary to predict the chemical shift influence of a pair of factors. The simplest approach to making a prediction for a give data point would be to find the cell which labels most closely resemble the factors values for the data point, and read the chemical shift value from the table. However, it is implicit in MINER (and SHIFTX) that the values in the table represent points on a curve, and for continuous input values that curve is smooth and continuous, thus some interpolation between cells is called for.

Splines

One problem that frequently arises in statistics and mathematical modeling is that of *interpolation*: given a set of points in an n-dimensional space that are presumed to be representative of a function $y = f(x_1,x_2,\ldots x_n)$, find way to estimate $f(x_n)$ for values of x which are not in the original set of points. One standard technique for this is the cubic spline, an interpolation formula that generates a curve passing through all the known points and continuous in both its first and second derivatives. Given a value x between the known points x_j and x_{j+1}, f(x) is given by (Press et al., 1992):

2-3) $$f(x) = Af(x_j) + Bf(x_{j+1}) + Cf''(x_j) + Df''(x_{j+1})$$

$$A \equiv \frac{x_{j+1} - x}{x_{j+1} - x_j} \qquad B \equiv 1 - A$$

$$C \equiv \frac{1}{6}(A^3 - A)(x_{j+1} - x_j)^2 \qquad D \equiv \frac{1}{6}(B^3 - B)(x_{j+1} - x_j)^2$$

The double-primes above signal the second derivative of the function, which is computed by solving the system of equations given by:

$$2\text{-}4)\ \frac{x_j - x_{j-1}}{6} f''(x_{j-1}) + \frac{x_{j+1} - x_{j-1}}{3} f''(x_j) + \frac{x_{j+1} - x_j}{6} f''(x_{j+1}) = \frac{y_{j+1} - y_j}{x_{j+1} - x_j} - \frac{y_j - y_{j-1}}{x_j - x_{j-1}}$$

Solving this system requires that arbitrary second derivative values be chosen for the end points of the curve. In MINER, these values were set to zero. Cubic splines were used for interpolation for all continuous factors used for shift prediction: angles (phi, psi, and chi) and bond lengths. The spline computation required that an arbitrary second derivative value be chosen for the end points of the curve. In MINER, this value was chosen to be zero.

One additional complication arose due to MINER's use of angles, which are not only continuous but cyclical: 180° and -180° represent the same point in angle space, but are opposite edges of a MINER table. More to the point, 160° and -160° are quite close in angle space, and should exert equal influence on the interpolated curve in the neighbourhood of 180°, but only 160° is conceptually "nearby" in the usual spline formulation. This problem was dealt with by duplicating the data twice, trebling the axes and in effect making them run from -540° to +540°. Using this method, the extreme values of the curve (near ±180°) are influenced by all nearby points, greatly reducing distortion of the curve's boundaries. This approach also moves the arbitrarily assigned zero second derivatives away from the region of the curve used to make predictions.

Making Predictions

The steps required for making predictions vary according to the nature of the factors used to construct the table. In particular, the different styles of axes require different computational steps for different scenarios. The simplest case is the one-dimensional discrete table (for example, if the single factor being tested is "Amino Acid Type") that reduces prediction to a simple lookup. Only slightly more complicated is the table with two discrete axes; again, only a lookup is required.

The next most simple case is the "continuous angle" or "continuous value, no zero" axis (for example, "Phi Angle"), where a single cubic spline computation is made to interpolate along the two-dimensional curve (for angle factors, trebling the axis as outlined above).

The prediction procedure for tables with one discrete axis and one continuous axis (for example, ("Amino Acid Type" vs. "Phi Angle" as in Table 2-2) is easily extrapolated from the procedures for one-dimensional tables. All points are assigned a row according to the value of their discrete attribute, and then a cubic spline interpolation is performed along the selected row (again, trebling the axis where necessary).

The most complex cases are tables with two continuous factors, requiring the application of the cubic spline in two dimensions, which is done in two steps. First, one interpolation is made along the horizontal axis for each row in the table, yielding one "intermediate value" for each row in the table. These intermediate values are then treated as a two-dimensional curve along which another interpolation is made, yielding the value sought. The presence of a zero bin in any of the above tables complicates matters only slightly. The zero row/column effectively becomes a separate one-dimensional table that

is applied to the subset of data points that have a zero value in the appropriate attribute. It is important to note that when making spline calculations (e.g. along a row) a zero cell in that row is ignored. Hence a zero row or column is effectively a separate table.

One of the more intricate parts of the MINER algorithm involves handling cases where the data is too sparse to permit an ordinary prediction. This most often happens due to the cubic spline algorithm's requirement for three data points along the curve. It may also occur due to the data segmentation used to statistically validate MINER's hypersurfaces (discussed in more detail elsewhere). Another common cause is data points that are missing some attribute; for example, the first amino acid in a protein has no "preceding amino acid". In all these cases, the challenge is to find a "reasonable" value for the chemical shift.

For missing attributes along one-dimensional tables, the data point is assigned a value of the mean of the entire table (excluding a zero cell on a continuous axis if present). If only the zero cell is occupied, the data point is assigned the value from the zero cell. If the table is completely empty, the data point is assigned a value of zero. The one-dimensional cases can be readily generalized to the two-dimensional case. If only one attribute is available, the data point is assigned the mean of the identified row or column. If both attributes are missing, the data point is assigned the mean of the entire table. If the table, row, or column is empty, the data point is assigned zero. In practice, these sparse-data conditions happened only very rarely, and were usually indicative of erroneous data (for example, phi or psi angles in the forbidden region of the Ramachandran plot) rather than exceptionally sparse data.

Evaluation

Having now generated a table and used it to generate a shift prediction for each input data point, the next step is to evaluate how well the table performs. The term "result" will be used in this section to refer to the correlation or RMS deviation between a table's chemical shift predictions and the actual chemical shifts. The simplest approach would be to take all the input data points and simply compute the result for the "best guess + new prediction", but this is statistically undesirable since it might well result in tables that perform well on the input data set, but poorly on other, unseen data. In order to test the both the efficacy and generality of the tables, the input data was divided into two or more segments, with one segment used to generate the table, and the others used to evaluate its effectiveness. This procedure is known as two-fold cross validation. Several different approaches to find the table's score (the figure of merit used to compare it to other tables) were tried:

- best segment result only;
- worst segment result only;
- sum of all segment results; and
- standard deviation of segment results.

This last approach is predicated on the notion that a table that returns consistent results is more likely to be correct than one which performs very well on some segments and very poorly on others. In practice, the various scoring methods returned similar results, with

the "sum of all segment results" shown to have the best performance overall (that is, resulting in a collection of tables that showed the best performance).

An early version of MINER allowed a "no prediction possible" value for data points that could not be fit into tables due to missing attributes; such points were then not considered in any subsequent table evaluation. Under these conditions, MINER would make its job easier by finding combinations of factors that excluded as many points as possible; generally speaking, fewer points mean higher correlations and lower RMS errors. Once this "laziness" was detected, the program was modified such that a prediction of zero would be made for input points that for one reason or another did not "fit" into the table. That particular table could then only improve its score by improving the prediction of the smaller group of points that it could affect, lowering the relative value of factors and factor pairs that could improve the prediction value of only a small subset of the input points. Thus MINER valued generality over specialization.

Picking The Winner

Through each iteration of MINER, all single factors and combinations of two factors are used to generate tables that are then scored. At this point, the tables are considered in order of descending score. If a single-factor table is the highest scoring, it is accepted without reservation. However, tables combining two factors are subject to additional testing to verify their usefulness and to avoid tables that, by chance, happen to produce good results by combining factors that are best considered independently. Such conditions are tested for by comparing the table's results to the results of the two factors applied as independent one-dimensional tables. That is, a new set of predictions is made,

using each of the factors in the original table in succession. For example, if the original table is "Phi vs. Amino Acid", then a "Phi" table is generated from the original data (before the application of the original table), the predictions of which are used to create temporary "new best guesses". These data is then used to generate an "Amino Acid" table from which set of predictions are generated; this set is then scored the same way as the original table's output. Another result set is generated by applying the single factors in the opposite order, "Amino Acid" followed by "Phi". The results of the three procedures are compared. If either of the "two tables applied sequentially" score equals or exceeds the score of the original table, then the latter is discarded, and the next-highest scoring table is evaluated. This entire procedure is diagrammed in Figure 2-1.

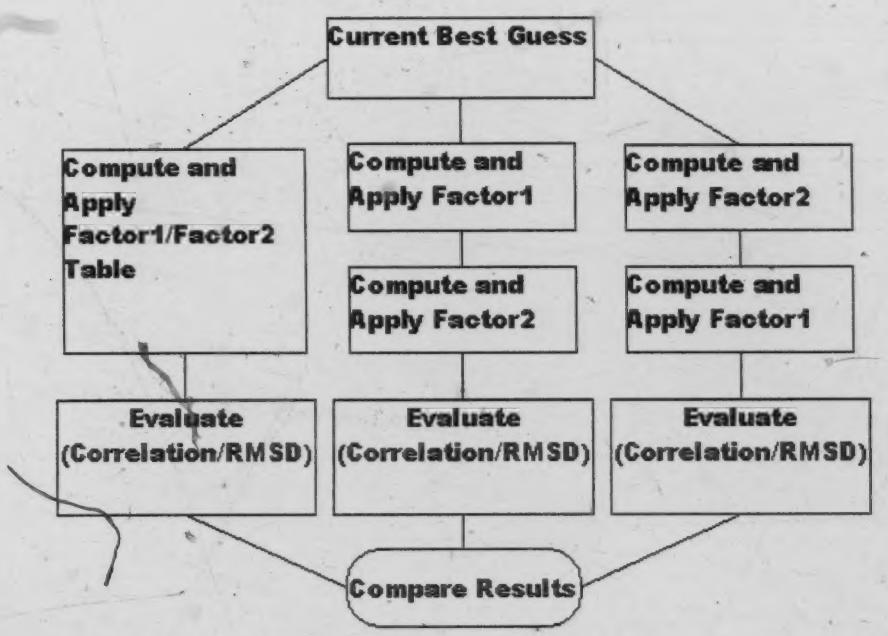


Figure 2-1: Comparing factors applied together to factors applied singly.

As mentioned, the purpose of the screening described above is to prevent the adoption of tables that, through statistical fluke, produce good results but do not reflect any physical reality or genuine relationship between the constituent factors. It can easily be seen that combining two independent factors into a single table will, at best, equal the results of the two factors applied independently. This best case scenario assumes that the two factors always interfere constructively and never destructively. In most cases, there will be destructive interference (where one factor's contribution is positive and the other is negative in sign) and the combined table must necessarily produce inferior results. This problem arises due to MINER's one-table-at-a-time operation. If MINER instead considered all possible combinations of tables, the problem would not arise, but such an evaluation would require prohibitive computational resources.

Once an optimally-performing table has been found, its predictions are added to the previous best guess to create a new best guess that will be used in subsequent iterations of the program. With this new best guess in hand, another set of tables is generated and evaluated. When the point is reached that no table significantly improves the results, the search is terminated and MINER outputs its findings and exits.

Numerical Optimization

MINER is a "greedy" algorithm, always taking the maximum possible step towards increased performance. The shortcoming of such algorithms is that large initial "steps in the wrong direction" may prove difficult or impossible to correct for in later steps. Also, situations may evolve where tables are mutually confounding, resulting in a

system that oscillates around the correct answer without ever finding it. Consider this contrived data set with four data points and two binary factors:

	Factor 1	Factor 2	Chemical Shift
Point A	2	2	12
Point B	2	2	12
Point C	4	2	18
Point D	4	4	24

Table 2-3: Contrived data set with 4 data points and two factors.

The "chemical shift" is simply triple the sum of the two factors, which can easily be seen to be uncorrelated. For purposes of this example, assume that factors 1 and 2 are each distillations of two non-independent factors such that they cannot be combined into a single table (which would make perfect predictions and spoil this example). Further assume that the "best guess" for each data point is zero. From these data MINER would generate two trial tables, shown in Table 2-4.

Factor 1	
Factor Value	Average Shift
2	12
4	21
New Best Guesses	
Point A	12
Point B	12
Point C	21
Point D	21
RMSD = 4.243	

Factor 2	
Factor Value	Average Shift
2	14
4	24
New Best Guesses	
Point A	14
Point B	14
Point C	14
Point D	24
RMSD = 4.9	

Table 2-4: Processing the data set of Table 2-3.

Factor 1 provides the best predictive value, and so would be used to generate the next generation of best guesses and associated error values, from which another pair of trial tables would be generated (Table 2-5).

Data Point	Actual Value	Best Guess	Error
Point A	12	12	0
Point B	12	12	0
Point C	18	21	-3
Point D	24	21	3

Factor 1	
Factor Value	Average Shift
2	0
4	0
New Best Guesses	
Point A	12
Point B	12
Point C	21
Point D	21
RMSD = 4.243	

Factor 2	
Factor Value	Average Shift
2	-1
4	3
New Best Guesses	
Point A	11
Point B	11
Point C	20
Point D	24
RMSD = 2.449	

Table 2-5: Second round of processing the data set of Table 2-3.

Even in this simple example, substantial errors remain in the predictions after the two input factors are applied. Further applications of the same process will improve the predictions, but the predictions will always oscillate around the correct values. The proximate cause of the problem can be seen by comparing the average shifts for the two values of factor 1 in the first set of trial tables. The actual contributions of the factors are 2 * 3 = 6 and 4 * 3 = 12, but the calculated values are 12 and 21 respectively. The errors are 6 and 9, exactly the average contributions by factor 2 over the sets of data points where factor 1 equals 2 and 4 respectively. The factor 1 table has "captured" contributions due to factor 2, and in general all tables will similarly capture contributions from all other tables that are calculated after them. This can be seen as a consequence of MINER's approach of searching for tables sequentially rather than simultaneously, and any attempt to correct the problem must allow some "feedback" from "late" tables to

"early" tables. During the development of MINER, two approaches to solving this problem were tried.

The "simple" optimization technique was applied after the second and subsequent tables were selected. The "best guess" resulting from the latest table was modified by removing the contribution of the first table and adding it back into the errors --- in effect, the influence of the first table was removed from the calculations. The first table was then regenerated from this modified data, and the best guess and error values updated appropriately. This procedure was then repeated on the second and subsequent tables until all tables had been processed, after which the performance of the collection of tables was evaluated and compared to the unprocessed tables. If performance improved, the optimized tables were adopted; if not, they were discarded and the unaltered set retained. This optimizing procedure was also repeated at the end of the search when no more useful tables could be found.

There is no strong theoretical basis for the efficacy of this procedure: its usefulness depends strongly on the statistical characteristics of the data. Nonetheless, it generally did result in substantial performance improvements in MINER, and had the further virtue of speed. However, it was desired to try an optimization technique with a firmer theoretical foundation, which led to the second optimization technique described below.

The fundamental insight in the discussion of "captured" contributions is that early tables absorb the average contributions of later tables, which are computed based on the residual errors "left over" after the early table has been applied. Designating each table as t1, t2, etc., using the subscript "11" to designate the set of points lying in the upper left

cell of table t1, "a1" to mean the actual contribution of the factors composing table t1, we have (assuming n tables):

2-5) $$t1_{11} = a1_{11} + \overline{t2_{11}} + \overline{t3_{11}} \ldots + \overline{tn_{11}}$$

As noted previously, t2, t3, and so forth have a complex dependence on t1, which is derived directly from and dependent only on the input data (assuming that it is not modified based on feedback from later tables). This second optimization technique explicitly recognizes the aggregate nature of t1 as the sum of an "actual value" and an offset, and manipulates the proportion of t1 assigned to each. The rightness of any given partition between the two can be tested by comparing the assigned offset value to the sum of the mean values of the other tables (for the same group of data points), and the offset adjusted until the two values are the same.

The fundamental idea is simple enough, but the implementation proved to be challenging. Unfortunately, it is not possible to manipulate a single offset in isolation; due to the different groupings of points in each table, changing a value in a single cell will generally affect many other cells, requiring that all cells in a given table be manipulated simultaneously. This necessitated the use of complex multi-dimensional optimization algorithms, requiring extensive computational resources in terms of both processor cycles and memory; a typical set of MINER-generated tables may include two thousand cells, making this problem equivalent to minimizing a function of two thousand variables. Even on a multi-gigahertz processor, using this style of optimization often required hours and days of processor time, versus seconds and minutes using the simple

optimization outlined above. In the end, the gains in accuracy and theoretical justification of this more complex optimization method were not justified by the time and resources required, and were reluctantly abandoned.

Testing

Testing MINER presented the typical challenge of a research tool: how can one be sure that an algorithm is correct and properly implemented if the "right" answers are not known? At one level, the output of early versions of MINER was validated by the fact that its broad conclusions agreed with earlier research, for instance that the backbone angles are the most important factor in predicting alpha carbon shifts. Further validation was provided by the fact that shift predictions based on MINER tables showed significant correlation with observed data, even data that MINER had not seen in reaching its conclusions. Nonetheless, rigorous testing was performed to ensure the reliability and accuracy of MINER's algorithms, and to confirm that the latter could perform the required tasks of deriving useful predictive tables from complex data.

In general, it was desired that any contrived test data set resemble the actual data in as many particulars as possible. In fact the test sets used were actual input sets, but with the true chemical shift value replaced with a value with a known (and usually complex) dependence on the input parameters. For example, in order to test MINER's ability to group interacting factors together, a test set using the product of the phi and psi angles as the chemical shift was presented to MINER. Any output other than a single table with phi and psi axes constituted a failure, as would any cells not approximately (some errors were inevitable due to the 40° wide divisions of the data) equal to the

product of the torsion angles. All the types of factors (discrete, continuous-wrap and continuous-zero) were tested individually and in weighted combinations to verify MINER's operation. The fundamental test cases are listed in Table 2-6 below.

Factors	Facility Tested
Bond Length	Handling of continuous-zero factors.
|Phi|	Handling of continuous-wrap factors.
Amino Acid	Handling of discrete factors.
Bond Length * Bond Length, Phi * Phi, Amino Acid * Amino Acid	Collapsing these pseudo-two-dimensional inputs into one-dimensional output tables.
(Bond Length + 1) * Amino Acid	Combination of continuous-zero and discrete factors. The +1 is included so that the zero column will still have a dependence on the amino acid code.
|Phi| * Amino Acid	Combination of continuous-wrap and discrete factors.
(Bond Length + 1) * |Phi|	Combination of continuous-zero and continuous-wrap factors. The +1 is included so that the zero column will still have a dependence on the phi angle.
Amino Acid * Previous Amino Acid	Combination of distinct discrete factors.
|Phi| * |Psi|	Combination of distinct continuous-wrap factors.
(H Bond Length + 1) * (O Bond Length + 1)	Combination of distinct continuous-zero factors.
|Phi| – |Psi|	Discrimination against combining factors that are actually independent.
5 * |Phi| – |Psi|	As above, but also testing that the most influential factor is extracted first.
|Phi| * |Psi| + |Phi| – |Psi|	Separation of factors that appear both alone and interacting with other factors.
|Phi| * |Psi| + |Phi| * Amino Acid	Dealing with factors that appear in multiple tables.

Table 2-6: Test cases for MINER. Vertical bars are used to indicate absolute values; these are often used with angular quantities to avoid discontinuities at sign changes. Amino acid types are converted into numeric values by using the ASCII code for the single-letter amino acid abbreviation.

Combinations, weight and unweighted, of all these test cases were also used to verify MINER's ability to separate multiple tables. Also tested were cases where the chemical shift was dependent upon three, four, or more interacting factors, to see if any

conclusions could be reached about how MINER performed with higher-dimensional data, and what loss of accuracy was entailed by estimating a high-dimension table with a succession of two-dimensional tables. No general principles emerged from this latter study, though there were some indication of a relationship between the entropy (in the information-theoretic sense of the term) of the input data and the accuracy of the two-dimensional estimate.

Discussion

A detailed discussion of the results of MINER will be postponed to the next chapter, where it will appear in the context of chemical shift prediction via SHIFTX. The discussion in this chapter will be more concerned with the development of the algorithms used in MINER and the refinement thereof.

As is usual in the development of any software, there was a steady progression from simplicity to complexity, driven by the quest for more accurate and precise results. Fortunately, there was a concomitant increase in computer power; over the time period during which MINER was developed, the clock speed of a processor of a given price increased by a factor of approximately six times, allowing ever more ambitious algorithms and processes to be tried. Nonetheless, the criteria for acceptance of any innovation were always the same: does it improve the results, lowering RMSD and increasing correlation, without decreasing the generality? Secondarily, does it yield results in a reasonable amount of time --- hours or days, rather than weeks? Seen in this light, the segmenting of the data into trial and test sets, and objective table evaluation criteria were vital in keeping development on track. Among the first subroutines written

for this program were "black box" functions for calculating correlation and RMSD with no inputs other than two sets of floating point numbers. Significantly, these subroutines were not changed in the course of development.

The other important element in the success of MINER was developing a standard set of test inputs to check that new changes did not introduce errors, subtle or otherwise, into the output. The usual process for testing software changes, comparing results the generated by the old version with the new version, was not always helpful with MINER. As noted previously, it was rare for a single table cell to change in isolation; more often, any change was reflected in large numbers of cells in many tables, making it difficult to assign significance to any given change. In cases like this, being able to use data that would generate a known output was invaluable.

One further challenge to the development and testing of MINER was data management. The input data sets all contained several thousand data points, each with more than twenty attributes. These data points also contained references to one another; for instance, each data point contained the amino acid type of the previous data point (which represented the previous amino acid in the chain). New attributes were added regularly, since one of the motivations for this project was to find novel factors for predicting chemical shifts. The tool of choice for managing these data sets was Microsoft Excel, which handled the task admirably in most cases. The limitation of Excel spreadsheets to 65,535 rows and 255 columns did not prove significant in this project, but should be borne in mind by anyone attempting to manage much larger data sets than used for MINER.

Conclusion

MINER started out as a tool for testing hypotheses about factors and factor combinations useful in predicting chemical shifts. Along the way it grew into a more general research tool useful for finding novel factors for shift prediction, as well as confirming the conclusions of past research and refining the findings thereof. Its value was further confirmed when some re-evaluation of past results required substantial alterations in the input data set; with the use of MINER, only a few hours were required to generate a corrected set of hypersurfaces for use with SHIFTX. It also greatly facilitated the answering of questions such as, "How good a prediction can be made if only glycine residues are considered?"

The most interesting further development of MINER would be to enhance the program to consider more than two factors at a time. This would greatly increase the time required to run the program, but is the step most likely to result in increased accuracy. As noted in the introduction to this chapter, MINER approximates a many-dimensional surface with a collection of two-dimensional tables; it is very likely that higher-dimensional estimators would make for a better approximation. One development that makes this improvement more likely to be practical is the now-widespread availability of processors incorporating SIMD (single instruction, multiple data) instruction sets (Intel, 2003) that greatly accelerate the array and matrix operations that make up much of MINER's computational requirements. In the meantime, however, two-dimensional tables have proven to be quite capable, as will be seen in the next chapter.

References

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Research*, **28,** 235-242.

Gusfield, D. (1997) *Algorithms on Strings, Trees, and Sequences*. Cambridge University Press, Cambridge.

Haykin, S. (1994) *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, Upper Saddle River, NJ

Intel Corporation (2003*) IA-32 Intel Architecture Software Developer's Manual Volume 2: Instruction Set Reference*. Intel Corporation, Denver, CO.

Langley, P. (1996) *Elements of Machine Learning*. Morgan Kaufmann Publishers, Inc San Francisco, CA.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wuthrich, K. (1998) *J. Biomol. NMR*, **12**, 1-23.

Press W, Teukolsky S, Vetterling W, Flannery B. (1992) *Numerical Recipes In C 2nd ed.* Cambridge University Press, Cambridge.

Rumelhart, D.E. and McClelland, J.L. (1986) *Parallel Distributed Processing*, The MIT Press, Cambridge, MA

Scheaffer R, McClave J. (1995) *Probability and Statistics for Engineers 4th ed.* Duxbury Press, Belmont, CA.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, 1,217-236.

Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR,* **6**, 135-40.

Chapter 3: SHIFTX*

Introduction

Chemical shifts are the "mileposts" of NMR spectroscopy. Not only are they important as spectral markers, but their dependency on multiple electronic and geometric factors means that chemical shifts can potentially provide a rich source of structural information. However, these multiple dependencies make both the interpretation and accurate prediction of chemical shifts exceedingly difficult – particularly for large molecules such as proteins. Fortunately, over the past decade, significant progress in chemical shift prediction has been made, both through computational advances (Williamson and Asakura, 1997; Case, 1998; Case, 2000; Wishart and Case, 2001) and through the rapid expansion of biomolecular chemical shift databases (Seavey et al., 1991; Zhang et al., 2003).

Currently there are three main approaches for calculating protein chemical shifts from atomic coordinates: 1) quantum mechanical, 2) classical, and 3) empirical. Quantum mechanical (QM) approaches employing density functional theory (DFT) have been used to very accurately calculate ^{1}H, ^{13}C and ^{15}N shifts for selected classes of residues in proteins (deDios et al., 1993; Le et al., 1995; Xu and Case, 2001, 2002). Classical approaches, which employ simplified or empirical equations derived from classical physics and experimental data, have been used to accurately calculate ^{1}H shifts for quite some time (Wagner et al., 1983; Dalgarno et al., 1983; Osapay and Case,

---

* A modified version of this chapter appeared as:
Neal S., Nip A.M., Zhang H. and Wishart, D.S. (2003) *J. Biomol. NMR*, **26**, 215-40.

1991,1994; Wishart et al., 1991; Herranz et al., 1992; Williamson et al., 1992). Empirical approaches, which rely on chemical shift "hypersurfaces" calculated from databases of observed chemical shifts, are capable of rapid, but only modestly accurate calculation of ^{1}H, ^{13}C, and ^{15}N shifts (Spera and Bax, 1991; Le and Oldfield, 1994; Beger and Bolton, 1997; Wishart and Nip, 1998; Iwadate et al., 1999). These hypersurfaces relate chemical shifts to various empirical parameters (backbone angles, nearest neighbors, side chain angles, secondary structure, etc.). Pre-calculated chemical shift hyper-surfaces are also used in QM approaches to greatly accelerate the speed of their calculations (Xu and Case, 2001, 2002; Le et al., 1995).

As of yet, none of the three approaches has developed to a stage where it can offer a rapid, accurate method to calculate all (i.e. side chain and backbone ^{1}H, ^{13}C and ^{15}N) chemical shifts for all residues under all conditions (diamagnetic and paramagnetic). Ideally, if one could combine the speed and predictive breadth of the empirical approaches with the accuracy of the classical or QM approaches, then it might be possible to achieve this goal. Here we describe a hybrid predictive method that attempts to combine the empirical hypersurface approach with the classical approach to accurately and rapidly calculate essentially all diamagnetic shifts for all 20 amino acid residues. The program, called SHIFTX, takes as input a protein structure in Protein Data Bank format, and predicts the diamagnetic ^{1}H, ^{13}C and ^{15}N chemical shifts of the protein's backbone and side chain atoms. Tests indicate that SHIFTX is rapid (about 1 second on a 2.2 GHz Pentium IV CPU for a complete shift calculation of 100 residues) and accurate. Overall, the program was able to attain a correlation coefficient (r) between observed and calculated shifts of 0.911 ($^{1}H\alpha$), 0.980 ($^{13}C\alpha$), 0.996 ($^{13}C\beta$),

0.863 (^{13}CO), 0.909 (^{15}N), 0.741 (^{1}HN), and 0.907 (side chain ^{1}H) with an RMS error of 0.23, 0.98, 1.10, 1.16, 2.43, 0.49, and 0.30 ppm respectively; these results are detailed in Table 3-5. We further show that the agreement between observed and SHIFTX calculated chemical shifts can be an extremely sensitive measure of the accuracy and precision of protein structures. We believe SHIFTX could serve as a valuable tool for refining and assessing protein structures, for validating and adjusting chemical shift assignments (Zhang et al., 2003) and potentially for generating 3D protein structures using only chemical shift information (Wishart and Case, 2001). A complete description of SHIFTX, its performance, applications and limitations follows.

Methods

SHIFTX employs a hybrid predictive protocol that uses a combination of classical equations and empirical "hypersurfaces" to calculate chemical shifts from atomic coordinate data. Classical equations are used to calculate the effects of well-characterized physical phenomena such as ring currents, H-bonds and electric field effects. The chemical shift hypersurfaces (described in more detail below) are derived from observed data and are fundamentally statistical in nature. These hypersurfaces serve as a simple method for capturing complex, nonlinear, and multi-parametric interactions that do not lend themselves to simple analytical expressions. A SHIFTX chemical shift calculation is therefore the sum of several components:

$$\delta_{calc} = \delta_{coil} + \delta_{RC} + \delta_{EF} + \delta_{HB} + \delta_{HS} \quad \text{3-1)}$$

where δ_{coil} is the random coil ^{1}H, ^{13}C or ^{15}N chemical shift (relative to DSS) of the amino acid as given by Wishart et al. (1995b), δ_{RC} is the ring current shift, δ_{EF} is the electric field contribution, δ_{HB} is the hydrogen bond contribution, and δ_{HS} is the contribution from the chemical shift hypersurfaces for the nucleus of interest (primarily the backbone dihedral angles). A SHIFTX prediction is composed of four phases:

1) reading the PDB file;
2) checking and adding hydrogen atoms (if necessary);
3) calculating the classical contributions (ring currents, electric field effects, etc.); and
4) calculating the chemical shift hypersurface contributions and summing them with the results of phase 3.

These calculations are performed for essentially all atoms that can yield measurable chemical shifts. A more detailed description of each of the four phases, including the specific formulae, criteria and protocols is given below.

File Input

SHIFTX reads standard PDB files using file I/O methods originally developed for VADAR (Wishart et al., 1994). The program will read in a specified chain from a multi-chain file (defaulting to the first chain if another is not specified) and ignores non-standard amino acids, non-water heteroatoms and ligands (heme rings, metals,

etc.). Conditions that may impact the accuracy of predictions (missing atoms, numbering irregularities, and chain breaks) are noted in the program output. The I/O portion of the program also loads the amino acid sequence, determines nearest neighbors, calculates dihedral angles, secondary structures, H-bond partners, salt bridges and charge pairs. These parameters are all used in evaluating the chemical shift hypersurfaces.

Hydrogen Placement

SHIFTX initially determines if there are hydrogen atom coordinates provided in the PDB file. If not, the position of HN atoms is calculated using the plane formed by the N, CA, and CO_{n-1} atoms. The proton is placed in this plane 0.86 Å from the N atom such that the angle formed by the H-N bond and the N-CO_{n-1} bond is 118.9 degrees. The HA atom (for non-glycine residues) is placed 1.0 Å from CA, such that it fills the 120 degree tetrahedron formed by CA, CB, N, and CO. For glycine, the two alpha hydrogens are placed by rotating the vector formed by CA and N around the vector formed by CA and C by +120 and -120 degrees. All other hydrogens are added using the program REDUCE (Word et al., 1999; http://kinemage.biochem.duke.edu.) which has been incorporated into the SHIFTX web server.

Ring Current Effects

The presence of aromatic rings and their associated ring currents can have a profound effect on the chemical shifts of nearby nuclei. As Osapay and Case of have shown (1991) the effects of these currents are best calculated using the semi-classical methods of Haigh and Mallion (1980). Our findings indicate that all hydrogen nuclei

can be affected by ring currents, as can the CA, CB, CO, and N atoms. In calculating the ring current contributions for a given protein, SHIFTX first generates a list of susceptible atoms and a list of rings, and then calculates the influence of each ring on each such atom. This influence is the product of a geometrical factor **G**, a target-specific constant **F**, and a ring-specific intensity **I** (see Table 3-1). The values of the latter two constants were determined through parameter fitting.

Residue Type	# of Ring Members	Intensity Factor	Ordered List of Members
Phe	6	1.05	CG CD2 CE2 CZ CE1 CD1
Tyr	6	0.92	CG CD2 CE2 CZ CE1 CD1
Trp	6	1.04	CD2 CE3 CZ3 CH2 CZ2 CE2
Trp	5	0.90	CG CD2 CE2 NE1 CD1
His	5	0.43	CG ND1 CE1 NE2 CD2

Table 3-1: Residue types containing aromatic rings, the number of ring members, and the associated "intensity factor". Note that Trp has two rings.

For each ring, a normal is computed from the cross-product of two vectors originating from the first ring member and extending to the second and last ring members, respectively. The ring is deemed to lie in a plane perpendicular to this normal. Next, the projection of the target atom onto the ring plane is found; this is designated point **O** for purposes of the following calculations. Finally, the areas of a series of triangles are computed, with the vertices of each triangle being adjacent points on the ring and the point **O.** Because each pair of adjacent ring atoms is considered (including the first and last atoms), there will be one triangle for each ring member. These areas are *algebraic* and may be positive or negative. For instance, consider vectors $\mathbf{R}_i$ and $\mathbf{R}_j$ from **O** to the i^{th} and j^{th} ring members. The area of the triangle formed

by these three points is negative if the cross product $\mathbf{R_i} \times \mathbf{R_j}$ is parallel to the ring normal, and positive if it is anti-parallel. The area of each triangle is then multiplied by a distance factor, given by

3-2) $$d_{ij} = \frac{1}{r_i^3} + \frac{1}{r_j^3}$$

where r_i = length of $\mathbf{R_i}$ and r_j = length of $\mathbf{R_j}$ Thus the geometrical factor **G** is given by:

3-3) $$G = \sum_{\substack{ring \\ members}} d_{ik} area_{ijO}$$

G, the ring-specific intensity **I**, and the target nucleus factor **F** are multiplied together to yield the total effect on the target nucleus' chemical shift due to the given ring:

3-4) $$\delta_{RC} = GIF$$

The ring-specific intensity factor (**I**) and target nucleus factors (**F**) were determined empirically using a simple grid search optimization protocol (step size of 0.01) on the training database. The residue-specific least square values (for **I** and **F**) suggested by Osapay and Case (1991) were used as starting values. Note that in this formulation, **F** corresponds to Osapay and Case's parameter B but with the implicit assumption that **F** will vary for different target nuclei (^{15}N, ^{13}C, ^{1}HN, etc.) as a result of their different shielding or differing electron cloud "mobility", whereas Osapay and Case's B applies only to protons. Initially, the ring-current optimization was performed only on the HA

shifts wherein all five ring-specific intensity parameters and the target nucleus factor (a total of six numbers) were allowed to vary. The resulting ring-specific intensity factors (**I**) are shown in Table 3-1 and were found to be quite similar to those reported by Case and Osapay. The resulting value for **F** (5.13×10^{-6}) also corresponds closely to the B value of 5.45×10^{-6} determined by Case and Osapay (1991). The ring current optimization process was then repeated for other target nuclei by holding the **I** values constant and allowing **F** to vary. Note that for heavy nuclei (^{15}N and ^{13}C) the grid step was changed to 0.10. The resulting target nucleus factors ($\times 10^{6}$) were: 7.06 for HN, 5.13 for HA and all other hydrogen atoms, 1.50 for CA and 1.00 for CB, CO and N. These variations in **F** seem to reflect the susceptibility of different nuclei to ring current effects, with amide protons being most susceptible and CB, CO, and N being least susceptible.

Electric Field Effects

SHIFTX uses the method of Buckingham (1960) to calculate the effects of electrostatic fields on chemical shifts. The shifts of alpha carbons and all hydrogens ("target" atoms) are subject to electrostatic effects; these effects may be caused by CO, O, OD_n, OE_n, or N atoms ("source" atoms). Apart from the types and coordinates of the source and target atoms, this calculation also requires the coordinates of the target's "partner" atom, to which it is bonded. A list of all the target atoms and their partners is available at the SHIFTX website. All sources influence all targets within 3.0 Å, with the following exceptions: 1) no source atom influence targets on its own or adjacent residues; 2) O (carbonyl oxygen) atoms do not influence HN (amide hydrogen) atoms

and 3) solvent (i.e. water) atoms do not act as sources. Each source atom has an associated partial charge, i.e. -0.9612×10^{-10} esu for O, OD_n and OE_n atoms, 1.3937×10^{-10} esu for C atoms and 0.7209×10^{-10} esu for N atoms. Given this information, the effect on the shift of each target by each source can be calculated as:

3-5) $$\delta_{EF} = \frac{1 \times 10^{22} q \varepsilon \cos\theta}{d^2}$$

where: $\varepsilon = 1 \times 10^{-12}$, q = source charge (in esu, see above), θ = angle formed by source-target-partner and d = distance from source to target (Å). The total effect on each target atom is the sum of the effects of each source atom.

Hydrogen Bond Effects

While it might be expected that electrostatic effects should account for most of the chemical shift perturbations brought on by nearby polar or charged atoms, we found that the explicit inclusion of hydrogen bond effects improved the overall performance in SHIFTX. The methods used to determine the presence of hydrogen bonds are modeled after those of VADAR (Wishart et al., 1994) with possible hydrogen bond donors being amide and alpha hydrogens (H and HA). The acceptors may be the carbonyl oxygens on the backbone (O), side chain oxygens (OD_n, OE_n, OG_n, OH_n), or oxygen atoms from water in the solvent. SHIFTX compiles lists of possible donors and acceptors, and considers the possible existence of a hydrogen bond between each donor-acceptor pair.

For a bond to exist, the donor and acceptor must be on different residues, and if the acceptor is a solvent oxygen, the donor must not be an HA. Also, the oxygen-hydrogen separation must be less than an empirically determined distance; 3.50 Å for

HN's and 2.77 Å for HA's. Bond geometry is also considered; specifically, the angle between the N-H bond vector and the C=O bond vector must be 90 degrees or more, computed with the vectors translated such that the C and N occupy the same point (Kabsch and Sander, 1983).

Having applied these rules to each donor-acceptor pair, SHIFTX then sorts the list of possible bonds by the O--H separation distance, shortest to longest. The list is then processed so that only the single "strongest" hydrogen bond is identified for each donor-acceptor pair. More specifically, the first bond on the list is deemed to exist, and to preclude the existence of any bonds involving the same donor or receptor; any such bonds are removed from the list. SHIFTX then moves on to the next bond on the now-culled list, and similarly removes any bond made redundant, repeating this procedure until the end of the list is reached. Tests conducted on a large portion of the training data revealed that the inclusion of multiple acceptors (i.e. bifurcated or trifurcated H-bonds) in the calculation of hydrogen bond effects actually diminished the overall performance of SHIFTX. This finding is consistent with the idea that hydrogen bonds are pseudo-sigma (i.e. covalent) bonds involving partial transfer of the hydrogen donor's single electron to its acceptor atom (Baker and Hubbard, 1984).

The formula developed by Wagner et al. (1983) and Wishart et al. (1991) for calculating the influence of hydrogen bonds on HA and HN shifts was adopted for this work. These workers found that an r^{-3} dependency (the reciprocal of the cube of the distance between donor and acceptor) was most consistent with their experimental data. We optimized their δ_{HB} parameters for amide hydrogens through a simple grid search

of the two variable parameters (step size of 0.01) using our much larger training database. For amide protons, the best fit formula accounting for δ_{HB} shifts is given by:

3–6) $$\delta_{HB} = \frac{0.75}{r^3} - 0.99$$

where: r = the hyrodgen-oxygen separation in Å. The above formula is valid for hydrogen bond lengths between 1.5 – 3.5 Å. While the vast majority (>95%) of hydrogen bonds involve amide protons, the work of Wagner et al. (1983), Wishart et al. (1991) and Derewenda et al. (1995) clearly shows that alpha protons can also be involved in hydrogen bonding and that this bonding can influence their chemical shifts. Consequently we included a second equation in SHIFTX to account for HA atoms which acted as hydrogen bond donors. For these hydrogen bonds a pseudo-sigmoidal potential was found to work best where the H-bond shift (δ_{HB}) is assumed to be constant between 2.61 – 2.77 Å (long H-bonds) and 2.00 - 2.27 Å (short bonds). For H-bonds between 2.61 – 2.27 Å the hydrogen bond contribution to the HA chemical shift is given by:

3–7) $$\delta_{HB} = \frac{15.69}{r^3} - 0.67$$

If a glycine residue happens to have H-bonds on both its HA's, the resulting contribution to the δ_{HB} chemical shift is the mean of the individual δ_{HB} shifts.

Empirical Chemical Shift Hypersurfaces

Previous studies (for example, Spera and Bax, 1991; Le and Oldfield, 1994; Wishart and Nip, 1998) have shown the utility of torsion angle chemical shift hypersurfaces in predicting chemical shifts; the essential idea is to use to a residue's backbone angles as parameters to a lookup table which returns a chemical shift value corresponding to those parameters. The empirical lookup tables described herein are an extension of the same idea to parameters other than local backbone torsion angles. To develop these tables or hypersurfaces, a special data-mining program, called MINER, was created to search through a large set of protein structural parameters thought to be important for chemical shift determination. MINER's task was to find which of those parameters or pairs of parameters was most effective in calculating atom-specific chemical shifts. "Effective" in this context meant minimizing root-mean-squared errors or maximizing correlation between the predicted and observed shifts; in practice, the two metrics produced essentially the same result.

In operation, MINER is supplied with a database in which each row is a residue, and each column is a physical parameter of that residue; the parameters supplied are listed in Table 3-2. In addition, each record contains the observed chemical shift for the nucleus of interest, and a "best guess" at that shift, the latter consisting of the sum of the classically calculable shift contributions (random coil, electrostatic, ring current, and hydrogen bond factors) for that nucleus. The input database is further divided into a "training set" and one or more "test sets"; any predictions generated from the training set were evaluated against the test set to prevent overfitting.

Factor	Previous Residue	Next Residue	Values/Discretization
First Residue	Y	N	True/False
Amino Acid Type	Y	Y	1 of 20
Phi Angle	Y	Y	2 (of 18) 40°-wide bins centered every 20° from 0-340°
Psi Angle	Y	Y	2 (of 18) 40°-wide bins centered every 20° from 0-340°
Chi Angle	Y	Y	One of 0-120°, 120-240°, or 240-360°
Chi2 Angle	N	N	One of 0-120°, 120-240°, or 240-360°
Secondary Structure	Y	Y	One of Coil, Helix, or Beta Sheet
Length of HA H-Bond	Y	Y	0, or one of 20 equal-length bins
Length of HA2 H-Bond	Y	Y	0, or one of 20 equal-length bins
Length of HN H-Bond	Y	Y	0, or one of 20 equal-length bins
Length of O H-Bond	Y	Y	0, or one of 20 equal-length bins
Disulfide Bond	Y	Y	True/False
HA Hydrogen Bond	N	N	True/False
HA2 Hydrogen Bond	N	N	True/False
HN Hydrogen Bond	N	N	True/False
O Hydrogen Bond	N	N	True/False
Hydrogen Bond Status	N	N	Concatenation of previous four values

Table 3-2: Physical or Derived Property Parameters Used By MINER.

MINER's fundamental task is building chemical shift hypersurfaces (or 2D matrices), with vertical and horizontal axes each corresponding to one of the pre-calculated protein structural or sequence parameters. Each residue in the data set was assigned to one or more cells in the table in accordance with the table's axes and the residue's parameter values. For example, if the table's axes were "phi" and "psi" respectively, each residue with phi and psi angles in the range -180° to -170° would be assigned to the top-left cell in the table; if instead the residue had a psi angle of -165°, it would instead be assigned to the second cell in the top row. The mean shift difference (the difference between the observed and predicted shift, $\Delta\delta$) of the group of residues in each cell was calculated, forming a table of $\Delta\delta$ values, which was then used to compute

a preliminary chemical shift hypersurface for each residue. In the case of continuous quantities (torsion angles and bond lengths), cubic splines were used to interpolate between table rows and columns; for discrete quantities (such as amino acid type) the surface was actually a simple lookup table. An example of a SHIFTX hypersurface is shown in Figure 3-1.

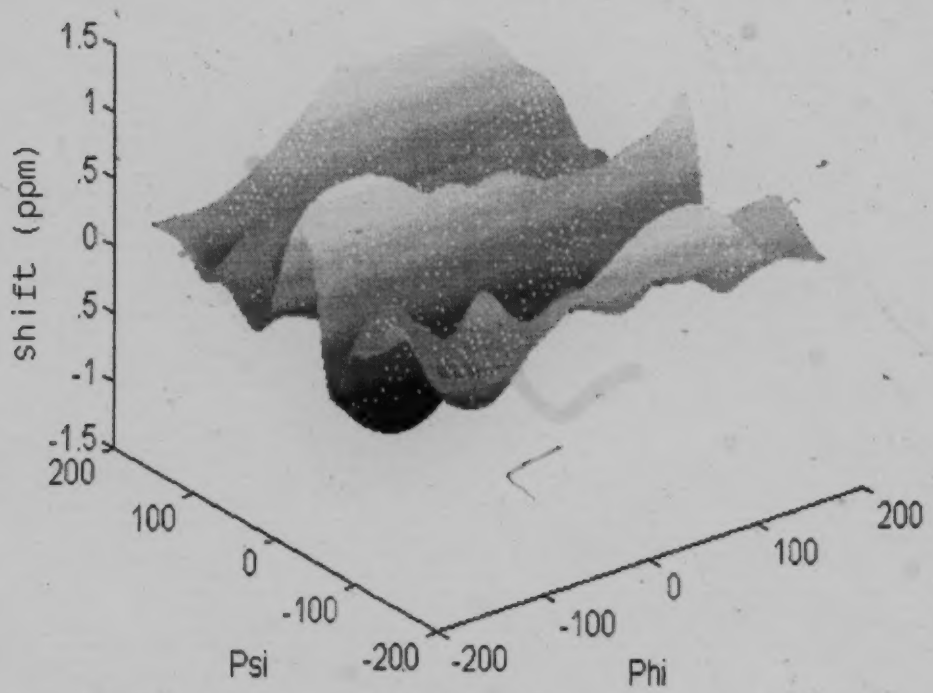


Figure 3-1: An example hypersurface.

Using the above technique, MINER generated and evaluated more than 400 possible hypersurfaces, noting those that were useful in predicting chemical shifts. Furthermore, because some of the structural parameters used in these hypersurface calculations were correlated to other parameters (both structural and chemical shift parameters), it was important to eliminate these co-dependencies. Consequently a hypersurface refinement

procedure was implemented in MINER to select and optimize the best set of hypersurfaces. Specifically, the refinement process involved five steps. First, for each possible pair of parameters, MINER constructed a table from the "training" set using the procedure outlined above, and computed $\Delta\delta$ values for all residues. Secondly, a trial "new best guess" for each residue in the test set was computed from the sum of the "best guess" and results of the previous step. The correlation of this "new best guess" to the observed chemical shifts was computed, as was the RMSD between the two. Third, after each possible pairing of parameters was evaluated in this way, MINER selected the pair yielding the highest correlation and/or lowest RMSD, and used the hypersurface thus generated to compute a new "best guess". Fourth, an iterative optimization procedure was applied to correct previous hypersurfaces, which may have "absorbed" some of the error that rightfully "belongs" to the new hypersurface. Finally, this procedure was repeated until no further improvement in RMSD or correlation was observed.

The above is a simplification of the actual procedure, which incorporated a number of features to preserve the statistical validity of the results, and selected against pairing factors that do not actually interact. Objections may be raised to the presence of the observed secondary structure as a parameter, when it is in fact derived from more fundamental parameters. Its inclusion here is a concession to the limitations of the two-dimensional surfaces which MINER generates. The secondary structure in this case operates as a mechanism by which MINER can select a subset of the records in the database and generate a different curve or predictive rule for each such subset.

Training and Testing Databases

Two databases of protein structures and the corresponding chemical shifts were used as input to MINER to generate the empirical constants, torsion angle surfaces, and lookup tables used by SHIFTX. One database was for backbone ^{1}H and heteronuclear (^{13}C and ^{15}N) shifts, the other was for ^{1}H side chain shifts. The database used to calibrate backbone ^{1}H, ^{13}C and ^{15}N shift predictions consisted of 37 diamagnetic proteins assembled from an extensive literature and BioMagResBank (Seavey et al.,1991) search. These proteins, their BMRB and PDB accession numbers as well as their resolution are shown in Table 3-3. In preparing this database every effort was made to find those proteins which had a) been reliably referenced (as indicated from the literature or BMRB data), b) spanned a variety of structural classes (all α, all β, mixed α/β), c) were well-structured, d) had no "shift-significant" ligands or paramagnetic moieties, and e) had high resolution X-ray structures (<2.1 Å). To prevent any instrumental or operator bias from creeping into the chemical shift calculations and the refinement process, assignment data was selected from a variety of labs. Additionally, several sets of protein shifts (especially ^{13}C and ^{15}N) were re-referenced to conform to IUPAC recommendations (Wishart et al., 1995b; Markley et al., 1998). Note that for backbone ^{1}H calculations and comparisons, all glycine ^{1}HA shifts were averaged and treated as a single shift because of the limited information on stereospecific assignments.

PDB ID	Protein Name	Resolution (Å)	BMRB Accession / Reference
2ALP	Alpha-lytic protease (L. enzymogenes)	1.70	(Cornilescu et al., 1999)
1GZI	Antifreeze Protein (Ocean pout)	1.80	(Wishart et al., 1997)
1A6K	Apomyoglobin (Sperm whale)	1.10	4061
1A2P	Barnase (B. amyloliquefaciens)	1.50	975
4ICB	Calbindin D9K, minor A form (Pig)	1.60	390
1CLL	Calmodulin (Drosophila)	1.70	547
1ROP	ColE1 Repressor protein (E. coli)	1.70	4072
1CEX	Cutinase (F. solani)	1.00	4101
3EZM	Cyanovirin-N (Nostoc ellipsosporum)	1.50	(Cornilescu et al., 1999)
2CPL	Cyclophilin-A (Human)	1.63	(Cornilescu et al., 1999)
1HCB	Carbonic anhydrase I (Human)	1.60	4022
1DMB	D-maltodextrin-binding protein (E. coli)	1.80	4354
1ICM	Fatty Acid Binding Protein (Rat)	1.20	(Wishart et al., 1997)
1HFC	Fibroblast Collagenase (Human)	1.56	4064
4FGF	Fibroblast Growth Factor (Human)	1.60	4091
1BKF	FK506 Binding Protein (Human)	1.60	4077
1HVR	HIV Protease (HIV)	1.80	(Cornilescu et al., 1999)
4I1B	Interleukin 1β (Human)	2.00	1061
3LZT	Lysozyme (Chicken)	0.92	4562
1LZ1	Lysozyme (Human)	1.50	(Wishart et al., 1997)
1ONC	P-30 Protein (Northern leopard frog)	1.70	4371
5PTI	Pancreatic trypsin inhibitor (Bovine)	1.00	46, 262, 485
1F3G	Phosphocarrier protein III glc (E. coli)	2.10	(Cornilescu et al., 1999)
1ACF	Profilin I (A. castellanii)	2.00	(Cornilescu et al., 1999)
1HKA	Pyrophosphokinase (E. coli)	1.50	4299
5P21	RAS P21 (Human)	1.35	(Wishart et al., 1997)
1RUV	Ribonuclease A (Bovine)	1.30	4031
2RN2	Ribonuclease H (E. coli)	1.48	(Wishart et al., 1997)
1RGE	Ribonuclease S (S. aureofaciens)	1.15	4259
2RNT	Ribonuclease T1 (Aspergillus oryzae)	1.80	(Wishart et al., 1997)
1SVN	Savinase (Bacillus lentus)	1.40	(Cornilescu et al., 1999)
1SNC	Staphylococcal Nuclease (S. aureus)	1.65	(Cornilescu et al., 1999)
1MKA	Thiol ester dehydrase (E. coli)	2.00	(Cornilescu et al., 1999)
2TRX	Thioredoxin (E. coli)	1.68	(Wishart et al., 1997)
1ERT	Thioredoxin - reduced (Human)	1.70	(Cornilescu et al., 1999)
1TOP	Troponin C (Chicken)	1.78	4401
1UBQ	Ubiquitin (Human)	1.80	(Cornilescu et al., 1999)

Table 3-3: PDB and BMRB files used to calibrate backbone and heteronuclear shift predictions

The other database, used to analyze the side chain hydrogens, was developed by searching the BMRB for chemical shift data for the nuclei of interest, and using that subset of the records for which there was a corresponding high-resolution X-ray crystal structure (2.0 Å or higher) in the Protein Data Bank (Berman et al., 2000). The selection criteria as described above was also employed for this database, however chemical shift referencing was not an issue for these shifts. The BMRB and PDB files used for the side chain ^{1}H shift database are shown in Table 3-4.

PDB ID	Resolution	BMRB Accession	PDB ID	Resolution	BMRB Accession
5PTI	1.00	48 1156 1179	1BM8	1.71	4254 4256
2SCP	2.00	4129	1PID	1.30	4266
1EPF	1.85	4162 4143	1AVS	1.75	4232
1R69	2.00	2539 195	1MJC	2.00	4296
1BRF	0.95	1991	1AIL	1.90	4317
1AAP	1.50	2024	1QST	1.70	4321
1HIP	2.00	2219	1EKG	1.80	4342
1FNF	2.00	2281	1ANF	1.67	4354
2PSP	1.95	2384	1EDH	2.00	4380
2MLT	2.00	245 606 36	1F2L	2.00	4397
4ICB	1.60	247 325	2WRP	1.65	916
1DOI	1.90	2472	1TGJ	2.00	4411
1F41	1.30	2476	1CTF	1.70	4429
1NOT	1.20	250 422 423	1AB1	0.89	4509
1IGD	1.10	2575	1AAZ	2.00	4459
1QHJ	1.90	2580	1RZL	1.60	4917
1IOB	2.00	2718 2719	1CY5	1.30	4661
3I[illegible]	2.00	280	1BEN	1.40	554
3CHY	1.66	3440 4472	1HOE	2.00	60 1816
1CKU	1.20	2999 3000	1UBQ	1.80	68
1DUZ	1.80	3078 3079	1BYF	2.00	4782
1ET1	0.90	3427 3449 1666	1BWI	1.80	1093
1PVA	1.65	144 3471 3472	1ZNI	1.50	1444 884 883
2ZTA	1.80	371	1FTG	2.00	5011
1RSY	1.90	4039 4041	1E65	1.85	1210
1BQU	2.00	4150	2OVO	1.50	1374
1ROP	1.70	4072	1NOA	1.50	1766
1CLL	1.70	4174	1PID	1.30	4266
1CBS	1.80	4186	1TN4	1.90	1553
1ORC	1.54	4207	1RCF	1.40	1580
1MOL	1.70	4222			

Table 3-4: PDB and BMRB files used to calibrate the side chain hydrogen shift prediction.

The resulting databases were divided into equal-sized test and training sets, with every other residue being assigned to the test set. As a further check against overfitting, all optimization steps were evaluated by testing their results against randomly-chosen samples from the databases. Several (usually twenty) such samples would be generated and evaluated (in terms of correlation or RMSD) against the optimized surfaces. Given two hypersurfaces yielding similar correlations or RMSDs, preference was given to the surface yielding the smallest variance of correlations/RMSDs among the random samples.

Results and Discussion

The correlation coefficients and RMS errors between the observed and calculated shifts for each atom type (^{1}H, ^{13}C, ^{15}N) as measured over all test proteins are listed in Tables 3-5 and 3-6. A more complete listing that breaks down the results in Tables 3-5 and 3-6 according to residue type and nucleus is provided at the SHIFTX website. These tables list the correlation between the observed values with the "best guess" values (incorporating the classical effects) and full SHIFTX predictions (which incorporate the hypersurfaces). To minimize the influence of probable misassignments and typographical errors in the input data, points for which the error (predicted minus observed) was greater than three standard deviations from the mean error were removed. Nuclei for which no observations were available (HH, HH11, HH12, HH21, and HH22) are not listed. Of the 24 side chain protons for which data was available, 14 had correlations greater than 0.85, and all but seven had correlations greater than 0.75. Five of the seven poor performers were nuclei for which stereoscopic labeling was

questionable; all seven were protons for which limited amounts of data (fewer than 350 shifts) was available.

Nucleus	Correlation (Physical Factors)	Correlation (All Factors)	RMSD (ppm)	Shift Range	RMSD / Shift Range	Number of Data Points
CA	0.897	0.980	0.98	31 (72 – 41)	0.032	4323
CB	0.990	0.996	1.10	59 (74 -15)	0.019	3281
CO	0.511	0.863	1.16	20 (185-165)	0.058	3135
N	0.626	0.909	2.43	40 (135 - 95)	0.061	4204
H	0.557	0.741	0.49	8 (12 - 4)	0.061	2993
HA	0.621	0.911	0.23	4.3 (6.3 - 2)	0.053	4437

Table 3-5: The correlation between the observed chemical shifts of the backbone atoms and SHIFTX predictions.

Nucleus	Correlation (Physical Factors)	SHIFTX Correlation (RMSD-ppm)	Number of Data Points	Notes
HB	0.9740	0.982 (0.21)	1172	For Ala, correlation is between the single observed value HB and the mean of the predicted values for HB1, HB2, and HB3
HB2	0.8788	0.924 (0.30)	2927	
HB3	0.8638	0.917 (0.31)	2854	
HD1	0.9906	0.992 (0.39)	987	For Leu and Ile, correlation is between the single observed value HD1 and the mean of the predicted values for HD11, HD12, HD13
HD2	0.9779	0.992 (0.33)	1216	For Leu, the correlation is between the single observed value HD2 and the mean of the predicted values for HD21, HD22, and HD23
HD21	0.2435	0.708 (0.26)	140	There appears to be some inconsistency in the nomenclature for HD21 and HD22. This was resolved by assigning the lower of the two to HD21 and the higher to HD22.
HD22	0.3278	0.801 (0.28)	140	See above
HD3	0.9277	0.944 (0.33)	471	
HE	0.9817	0.987 (0.46)	127	For Met, correlation is between the single observed valued HE and the mean of the predicted values for HE1, HE2, and HE3
HE1	0.6575	0.912 (0.51)	444	
HE2	0.9902	0.985 (0.35)	507	
HE21	0.3051	0.778 (0.24)	113	Inconsistency in the nomenclature for HE21 and HE22. This was resolved by assigning the lower to HE22 and the higher to HD21.
HE22	0.4374	0.743 (0.26)	113	See above
HE3	0.9926	0.994 (0.20)	260	
HG	0.5878	0.700 (0.26)	318	
HG1	0.6178	0.696 (0.20)	333	For Val, correlation is between the single observed valued HG1 and the mean of the predicted values for HG11, HG12, and HG13
HG12	0.3778	0.464 (0.37)	215	
HG13	0.4591	0.566 (0.35)	208	
HG2	0.9201	0.928 (0.25)	1925	For Ile, Val, and Thr, correlation is between the single observed valued HG2 and the mean of the predicted values for HG21,HG22,HG23
HG3	0.8027	0.846 (0.29)	1041	
HH2	0.5820	0.858 (0.19)	61	
HZ	0.4862	0.674 (0.40)	162	
HZ2	0.6329	0.851 (0.23)	60	
HZ3	0.4197	0.868 (0.18)	59	

Table 3-6: The correlation between the observed chemical shifts of the side chain protons and SHIFTX's predictions.

Hypersurfaces

The utility of the hypersurfaces in improving predictions is obvious when the "physical factors only" predictions are compared with those made using the hypersurfaces. The improvement is particularly notable in cases such as backbone ^{15}N shifts, where the correlation increased from 0.626 to 0.909 when the hypersurfaces were applied. As with all of the backbone nuclei, most of the improvement was due to the application of a hypersurface utilizing the backbone torsion angles. Other minor improvements in ^{15}N prediction were possible with the addition of hypersurfaces indexed by residue type/chi angle and predicted secondary structure/preceding residue; this latter captures the "nearest neighbour" effect noted by earlier workers (Wishart et al., 1995a; Braun et al., 1994).

Up to eleven hypersurfaces were generated for each nucleus. Space limitations preclude a detailed listing in this paper, but the complete tables are available on the SHIFTX website. To gain a better understanding of which factors (both physical and empirical hypersurfaces) most influenced backbone chemical shifts, we have tabulated and enumerated these effects in Table 3-7.

Factors	^{1}HA	^{1}HN	^{13}CA	^{13}CB	^{13}CO	^{15}N
Φ/Ψ	38.7		46.2	30.7		6.9
Ring Current	13.8	14.5	4.8		0.7	0.2
AA/Φ	11.7				13.4	
Ψ/O–$Bond_{i-1}$	10.8					
Electric Field	4.8	3.9	1.6			
Ψ_{i-1}/HN Bond		25.3				
H-Bond	1.2	16.7				
Φ_{i-1}/O-Bond		10.4				5.2
Φ/Φ_{i+1}		7.9	0.9			
Ψ/Ψ_{i-1}						29.2
χ_{i-1}/Ψ_{i-1}			4.3			15.0
AA/χ			12.8	14.3		13.6
AA_{i-1}/SS_{i-1}						11.9
Ψ/Φ_{i+1}			9.6			
Φ/HA1-Bond			5.0			
Ψ/SS				13.3		
Ψ/χ				12.0	32.1	
χ/Ψ_{i-1}					10.7	
Φ_{i+1}/Ψ_{i+1}					9.6	
SS/Ψ_{i-1}					5.3	

Table 3-7: Relative influences of various factors and hypersurfaces on secondary shifts. The values shown are average "percentage of influence" on secondary shift for all residue types; details of this calculation can be found below. The subscripts "i+1" and "i-1" indicate " the following residue" and "the preceding residue" respectively. AA corresponds to amino acid and SS corresponds to secondary structure. A blank cell indicates that the hypersurface or factor does not influence that particular nucleus.

In this table we have identified the most dominant physical factors (electrostatic, ring current, hydrogen bond, disulfide bond, etc.) along the most prominent hypersurfaces

and enumerated their percent contribution to the calculated chemical shifts for each backbone nucleus. To compute these values, the "total influence" for each shift prediction was defined to be the sum of the absolute values (in ppm) of each factor, and "relative influence" to be the ratio of the absolute value of each factor to the "total influence". These relative influences were averaged over all the relevant residues, and multiplied by one hundred to express them in percentage terms in Table 3-7. A more complete listing of these tables (broken down according to residue type and including all eleven hypersurfaces) is available at the SHIFTX website.

In all cases, the effects of the backbone torsion angles are dominant. For instance, looking at the HA predictions, four factors contribute almost 75% of the influence when all residue types are considered. It is worth noting that these factors and their relative contribution roughly correspond to the values proposed by Wishart and Case (2001) and are in reasonably good agreement with the results determined by Xu and Case (2001; 2002).

As might be expected from statistically condensing a complex geometrical phenomenon into a two-dimensional table, the torsion angle hypersurface for HA (as seen in Figure 3-1) is quite complex. This aspect of SHIFTX -- approximating an extremely complex function as a sum of simpler functions -- naturally raises the question of how independent these simple components are of one another. To investigate this question further, principal component analysis (PCA) was applied to a subset of the SHIFTX training/testing data. PCA is a robust statistical technique in correlation analysis that allows one to independently and unambiguously identify the most prominent effects or contributions to a given phenomenon or calculation. It also

allows certain unexpected co-dependency of certain parameters (i.e. hypersurface components or factors) to be identified. Shown in Table 3-8 is a principal component analysis of lysine HA chemical shift data taken from the SHIFTX training/testing database. This table shows the relative influence of the 14 SHIFTX parameters used in calculating lysine HA chemical shifts. All columns had some influence, although the influence of components 7-14 is quite small. The key point from this analysis is that the factor space did not collapse into a small number of columns, which would indicate a strong co-dependence among the input factors. That said, it is worth noting that 90% of the variance is accounted for by the first three principal components. These data are actually quite consistent with the results shown in Table 3-7, which show that most chemical shifts for most nuclei depend primarily on three or four physical factors or hypersurface contributions, though the most influential factors are different for different nuclei. The more minor contributions just add a few percentage points to the overall quality of the fit.

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6	Comp 7
Percent Variation	57.1393	27.4464	5.6123	4.7552	1.7781	0.8454	0.6477
Phi/Psi	0.9608	-0.2448	0.0132	0.0485	-0.1033	0.0177	-0.0493
Ring Current	0.2501	0.9673	0.0022	-0.0101	0.0307	0.0023	-0.0097
Amino Acid/Phi	0.0515	-0.0067	0.0082	0.0443	-0.0202	-0.1963	0.9161
Psi/O Bond-1	-0.0668	0.0316	0.5376	0.8076	-0.2238	0.0397	-0.0329
Electrostatic	0.0753	-0.0487	0.3712	0.0072	0.821	-0.3981	-0.088
Phi+1/HA1 Bond	0.0149	0.0024	-0.3048	0.2897	0.3687	0.2541	0.1404
Secondary Structure/Chi+1	0.0261	-0.0213	0.0381	0.0342	0.314	0.7626	0.0643
Secondary Structure/Psi+1	0.0103	-0.008	0.021	0.0148	0.1003	0.237	0.2784
Chi/HA2 Bond	0.0075	0.0128	0.0003	-0.0037	-0.0369	-0.1518	0.1847
Chi-1/ Secondary Structure-1	-0.001	0.0138	-0.0057	-0.0002	-0.0012	-0.1134	0.0193
Psi/Disulfide-1	-0.0203	0.0063	0.0041	-0.0053	-0.0357	-0.041	0.0293
Hydrogen Bond	-0.0032	0.0058	0.6914	-0.5078	-0.122	0.2469	0.1127
First Residue/ HA2 Bond	-0.0058	0.0004	0.0031	0.0072	-0.0411	0.0118	-0.007
HA1 Bond/ HA2 Bond	0	0	0.001	-0.0008	-0.0001	0	0.0002
	Comp 8	Comp 9	Comp 10	Comp 11	Comp 12	Comp 13	Comp 14
Percent Variation	0.5664	0.4222	0.3316	0.2047	0.14	0.1108	0
Phi/Psi	0.0216	0.0127	0.0031	0.0092	0.0173	0.0013	0
Ring Current	-0.0177	0.0017	-0.0096	-0.0119	0.003	0.0002	0
Amino Acid/Phi	-0.0951	-0.2363	0.0049	-0.2275	0.0203	-0.0075	-0.0001
Psi/O Bond-1	0.0107	0.0011	0.0071	0.0085	-0.0063	-0.0166	0.0001
Electrostatic	-0.1056	0.0109	-0.0261	-0.0173	0.0198	0.0346	-0.0002
Phi+1/HA1 Bond	0.7682	0.0023	-0.1052	0.0371	0.0061	-0.0078	0.0004
Secondary Structure/ Chi+1	-0.3924	-0.3332	0.1777	0.1035	0.0409	0.0543	0.0001
Secondary Structure/ Psi+1	-0.1834	0.9034	-0.0267	0.0483	-0.0454	-0.0234	0
Chi/HA2 Bond	-0.0157	-0.0749	-0.0619	0.942	-0.1832	0.1035	-0.0001
Chi-1/Secondary Structure-1	0.1647	0.085	0.969	0.0311	-0.0448	0.1013	0.0001
Psi/Disulfide-1	0.0163	0.0507	0.0228	0.1784	0.9793	0.0381	-0.0001
Hydrogen Bond	0.4162	-0.0119	-0.0386	0.0137	-0.0106	-0.0152	-0.0012
First Residue/ HA2 Bond	0.017	0.0346	-0.1047	-0.1139	-0.0181	0.986	0
HA1 Bond/ HA2 Bond	0.0002	0	-0.0001	0.0001	0	0	1

Table 3-8: Principal component analysis of the prediction factors for lysine HA shifts. Each column is a "synthetic" component created by PCA, and the values in the table are conceptually equivalent to the coefficients of a multiple regression. The top row shows the percentage of the variance accounted for by the given component.

Performance of ^1H Predictions

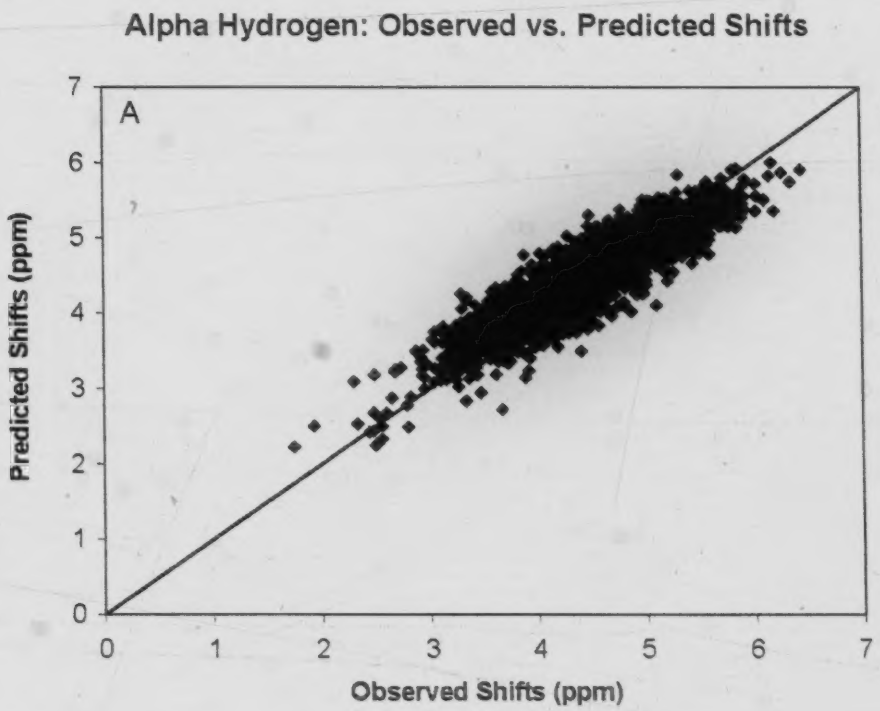
Overall, ^1H shift predictions are seen to be highly correlated with experimentally observed shifts, especially for the side chain ^1H's. As expected, the weakest correlation is obtained for amide protons. These results are highly consistent with previously reported ^1H shift prediction methods (Williamson et al., 1992; Osapay and Case, 1991; Herranz et al., 1992). For instance the HA performance reported here of 0.911 compares favorably to the value of 0.849 reported for Osapay and Case (1991), 0.747 reported for Asakura et al. (1992) and 0.730 reported for Herranz et al. (1992). Likewise the HN results of 0.741 reported here compares well to the value of 0.575 reported for Osapay and Case (1991) and 0.711 reported for Herranz et al (1992). Similarly, the correlation coefficient of 0.907 reported here for side chain ^1H's compare favorably to the value of 0.899 reported by Osapay and Case (1991).

To conduct a more controlled comparison, we used the publicly available programs from Williamson (TOTAL) and Case (SHIFTS) to predict the backbone ^1H shifts for a selection of five proteins (PDB accessions: 193L, 4ICB, 1POH, 5PTI, and 5TNC) covering 520 residues. We obtain correlations for HA predictions of 0.796 (TOTAL) and 0.742 (SHIFTS) versus 0.917 for SHIFTX. For HN predictions we get 0.548 (TOTAL) and 0.336 (SHIFTS) versus 0.756 for SHIFTX. Clearly SHIFTX does comparatively well in proton shift calculation, however, there is still room for improvement – particularly for amide protons. This suggests that either we still have an imperfect understanding of all the contributions that lead to protein ^1H chemical shift variation (especially HN shifts) or that the prediction methods have reached their limit

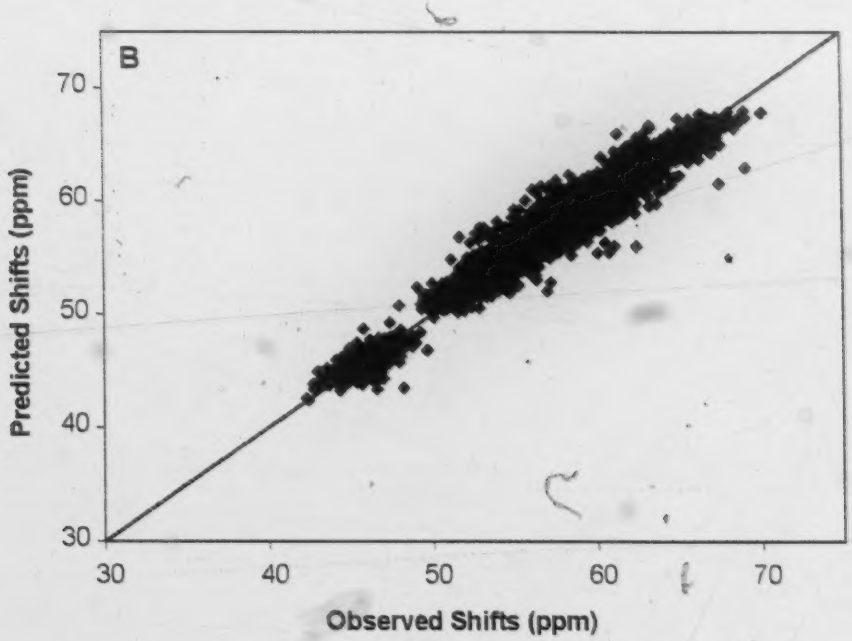
because of coordinate imprecision in the training or testing data. Our inclination is to believe it is more the latter than the former.

Performance of ^{13}C and ^{15}N Predictions

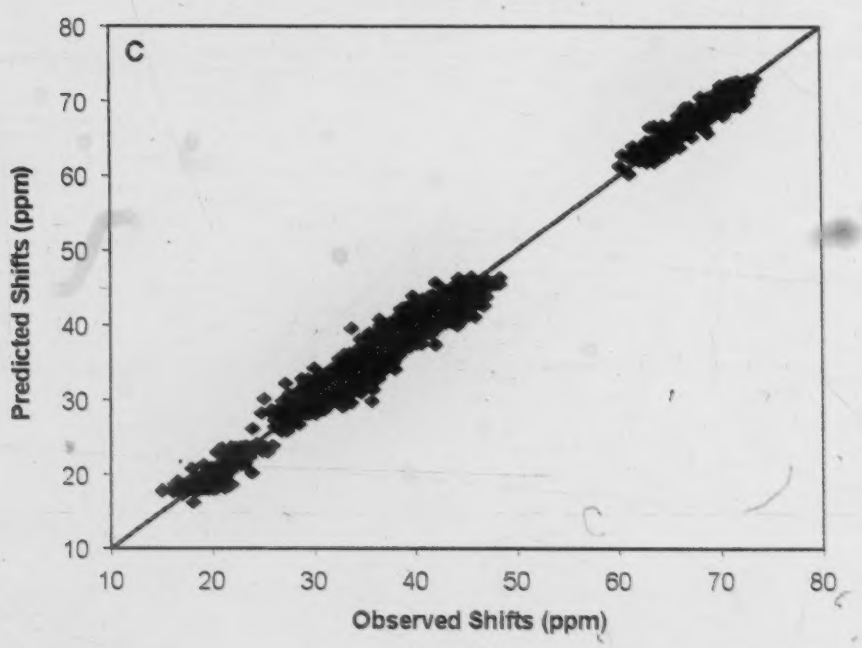
Figure 3-2 shows scatterplots illustrating the correlation coefficient (r) for ^{13}C and ^{15}N shift predictions based on the analysis of 4323 ^{13}CA, and 4204 ^{15}N shifts collected from 37 proteins with X-ray structures having a resolution less than 2.1 Å.

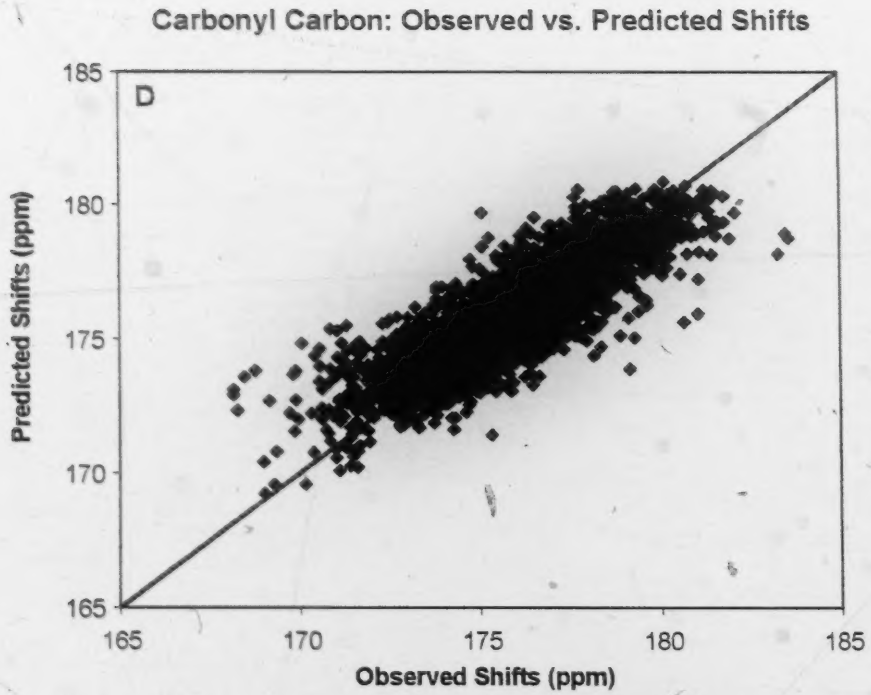


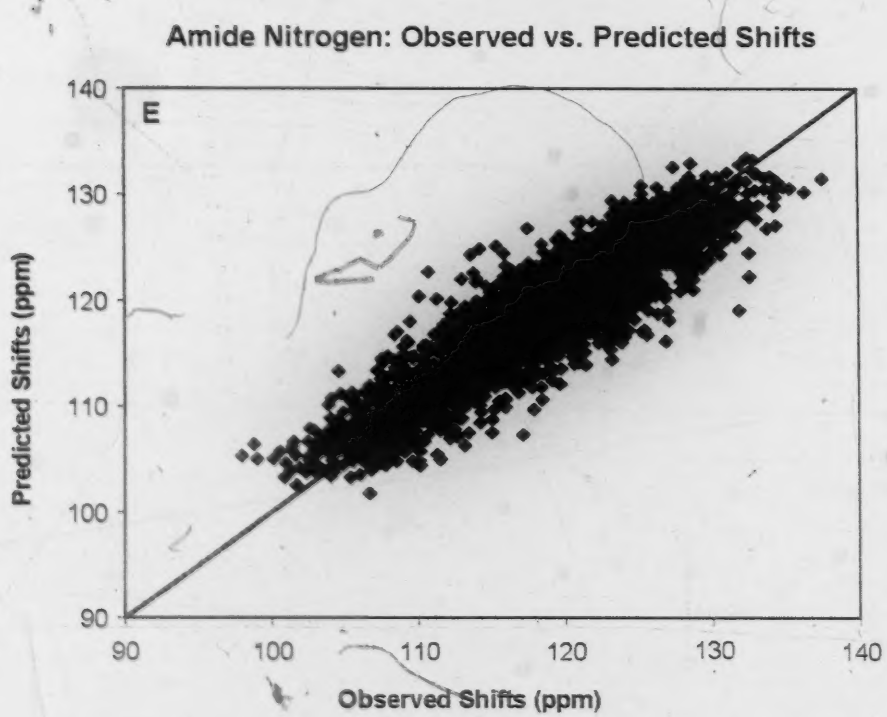
Alpha Carbon: Observed vs. Predicted Shifts



Beta Carbon: Observed vs. Predicted Shifts




Carbonyl Carbon: Observed vs. Predicted Shifts
D
Predicted Shifts (ppm)
Observed Shifts (ppm)
185
180
175
170
165
165
170
175
180
185


Amide Nitrogen: Observed vs. Predicted Shifts
E
Predicted Shifts (ppm)
Observed Shifts (ppm)
140
130
120
110
100
90
90
100
110
120
130
140

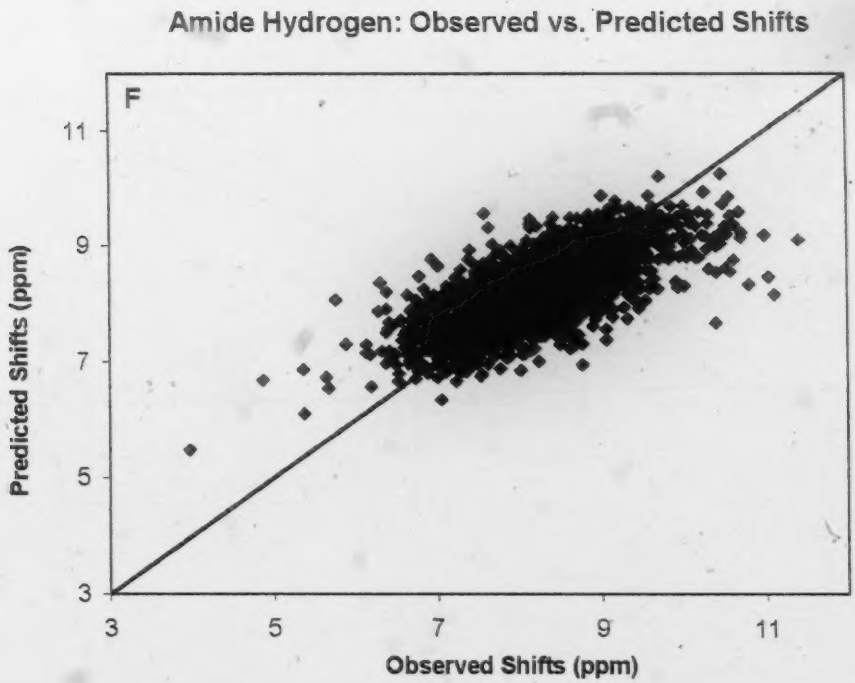


Figure 3-2: Scatterplots comparing observed vs. SHIFTX predicted shifts for backbone nuclei including A) $^{1}H\alpha$; B) $^{13}C\alpha$; C) $^{13}C\beta$; D) ^{13}CO; E) ^{15}N; and F) ^{1}HN. A line with a slope of one is drawn in each graph for comparison purposes.

Overall, the predictions are seen to be highly correlated with the range of experimentally observed shifts, especially for the ^{13}CB's. As expected, the weakest correlation is obtained for $^{13}C'$ shifts. Earlier studies on ^{13}C and ^{15}N shift prediction are somewhat limited as they only reported results from a small or well-defined subset of residue types. For instance, de Dios et al (1993) reported correlations of 0.97, 0.93 and 0.97 for ^{13}CA, ^{13}CB and ^{15}N shifts for ~20 alanine and valine residues in calmodulin and staphylococcal nuclease. Wishart and Nip (1998) report correlations of 0.97 and 0.80 for the complete set of ^{13}CA and ^{15}N shifts for calmodulin (148 residues) using empirically derived chemical shift hypersurfaces relating backbone dihedral angles to ^{13}C and ^{15}N shifts. An unpublished neural network approach (called PROSHIFT and

located at http://www.jens-meiler.de/proshift.html) has also been described, which claims to be able to predict ^{1}H, ^{13}C and ^{15}N shifts with RMSD errors of 0.22 pm, 0.98 and 2.08 ppm respectively. More recently Xu and Case (2001), reported correlations of 0.98, 0.99, 0.90 and 0.92 for ^{13}CA, ^{13}CB, ^{13}CO and ^{15}N shifts for residues (excluding His and Cys) from a set of 20 proteins found in well-defined helices and beta strands (which accounts for about 40% of all residues in proteins). While this manuscript was under review, Xu and Case (2002) published an extension to their earlier paper wherein they expanded their analysis to include residues from unstructured regions for the same 20 proteins (excluding cysteines as well as C and N terminal residues). They now report correlation coefficients of 0.97, 0.99, 0.83 and 0.90 for ^{13}CA, ^{13}CB, ^{13}CO and ^{15}N shifts with RMSD errors of 1.22, 1.31, 1.28 and 2.71 ppm respectively. Currently their method is able to predict ^{13}C and ^{15}N chemical shifts for about 89% of all residues in diamagnetic proteins.

As a comparison, the correlation coefficients we obtained for SHIFTX predictions over a set of 37 proteins were 0.980, 0.996, 0.863 and 0.909 for ^{13}CA, ^{13}CB, ^{13}CO and ^{15}N shifts with RMSD errors of 0.98, 1.10, 1.16 and 2.43 ppm respectively for all residues (100% coverage). Despite their differences in derivation and testing, overall the performance for both SHIFTX and SHIFTS (version 4.1, Xu and Case, 2002) for heteronuclear chemical shift calculation appears to be similar. It is possible that by combining SHIFTX and SHIFTS predictions together, one may be able to modestly improve the overall quality of heteronuclear chemical shift calculations. Efforts are underway in our laboratory to investigate this possibility.

The key point we would like to make here is that the ^{1}H, ^{13}C and ^{15}N shifts calculated by SHIFTX were for all residues, in all conformations and for all proteins in the test set. No prior geometrical optimization or energy minimization were performed on the input PDB file and the calculations for all ^{1}H, ^{13}C and ^{15}N shifts for each protein were done in less than 1 CPU second. We believe these performance characteristics are essential for using the program in practical situations pertaining to structure refinement, protein assignment, assignment validation and structure validation.

Secondary Validation

A frequent complaint about many predictive methods is that they work well on the training and test data, but fail miserably when tried on a "novel" set of previously unseen data. This is the all-too-familiar problem of over-training. As a further check on the validity and generality of SHIFTX we applied the program to predicting the ^{1}H, ^{13}C and/or ^{15}N chemical shifts of a set of comparable high resolution X-ray structures which were not included in the original SHIFTX test/training set. Ten such proteins (Table 3-9) were identified. The correlation and RMSDs between the experimental and SHIFTX predicted shifts for all backbone nuclei are shown in Table 3-10, along with the corresponding values for the original training data. The results for the novel proteins are clearly comparable to those for the training data, and in the case of the amide hydrogens, the correlation is actually better for the previously-unseen proteins. We believe that these results adequately demonstrate the generality of SHIFTX, which is to say that the methods and formulas used by SHIFTX to make predictions are valid outside of the proteins used to arrive at those methods and formulas.

Name	PDB file	Chain #	Resolution (Å)	BMRB Accession
Bucandin	1F94	A	0.97	5097
Heterogeneous Nuclear Ribonucleoprotein A1 (HNRP)	1L3K	A	1.10	4084
T Cell Signal Transduction Molecule	1D4T	A	1.10	5211
Plastocyanin	1PLC		1.33	4019
Human beta-defensin-2	1FD3	D	1.35	4642
Thermal hysteresis protein isoform yl-1	1EZG	B	1.40	5323
Histidine-Containing Phosphotransfer Domain of Arcb	2A0B		1.57	4857
Retinoic acid binding protein	1CBS		1.80	4186
Glur2 Ligand Binding Core	1M5E	A	1.46	5182
Iib Cellobiose	1IIB	B	1.80	4955

Table 3-9: Proteins used to validate SHIFTX performance on previously unseen data.

	Validation Set			Training Set		
Nucleus	Correlation	RMSD Error	Observed Shifts	Correlation	RMSD Error	Observed Shifts
HA	0.895	0.26	748	0.911	0.23	4437
H	0.746	0.52	1001	0.741	0.49	2993
N	0.901	2.53	904	0.909	2.43	4204
CA	0.979	1.02	866	0.980	0.98	4323
CB	0.996	1.10	778	0.996	1.10	3281
CO	0.856	1.17	758	0.863	1.16	3135

Table 3-10: The correlation, RMSD, and number of observed shifts for ten proteins not used in the training data, and the corresponding values for the training data.

Applications

Chemical shift calculations, if sufficiently accurate, can have a wide range of practical applications in protein NMR. These include 1) aiding chemical shift assignment; 2) chemical shift assignment validation; 3) chemical shift reference checking; 4) structure refinement; 5) structure evaluation; and 6) structure generation. For instance, Williamson et al. (1995) have shown how ^{1}H shift predictions based on the structure of the G domain of protein B1 were clearly able to identify two experimental misassignments. This paper also demonstrated that there is a good correlation between the accuracy of predicted ^{1}H shifts and the resolution of the corresponding structure. Subsequent studies by Kuszewski et al. (1995a; 1995b) and Osapay et al. (1994) showed that chemical shift refinement using ^{1}H, and then later, ^{13}C shift "calculators" could be used to improve the quality of NMR-derived protein structures.

Here we wish to demonstrate how SHIFTX can be used in three specific applications: a) chemical shift reference checking; b) chemical shift assignment validation; and c) structure evaluation. The first application concerns a key issue in heteronuclear NMR – that is the inconsistency of chemical shift referencing for ^{13}C and ^{15}N nuclei. This subject has been discussed extensively in a number of recent reviews (Wishart and Case, 2001; Wishart and Sykes, 1994). While clearly defined protocols do exist (Wishart et al., 1995b; Markley et al., 1998), it is estimated that nearly 25% of newly reported protein ^{13}C and ^{15}N shifts are improperly referenced (Zhang et al., 2003). A key issue is how to identify and correct these chemical shift referencing errors. Here we show that by using the 3D structure (either X-ray or NMR) of the

protein of interest and calculating its ^{13}C or ^{15}N shifts with SHIFTX it is relatively easy to detect and correct these referencing errors. Illustrated in Figure 3-3 is a plot of the observed versus SHIFTX-calculated ^{13}CA shifts for ribonuclease H (BMRB-1657; PDB 2RN2).

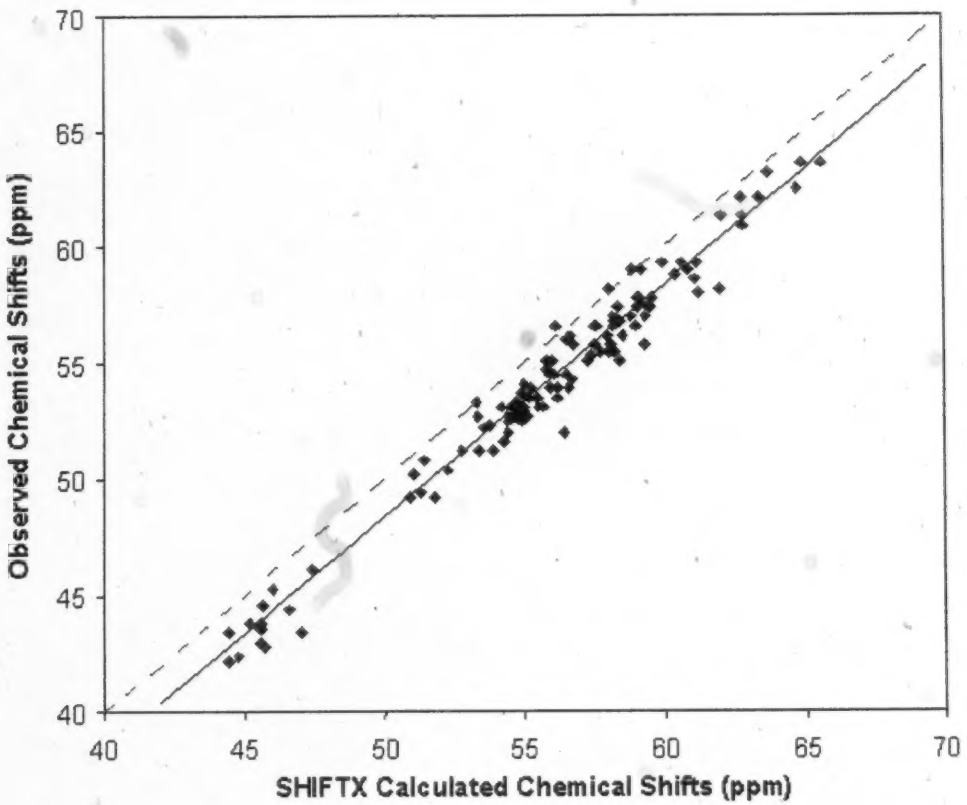


Figure 3-3: Plot of the observed vs. SHIFTX predicted $^{13}C\alpha$ shift of ribonuclease H (BMRB1657). A dashed line is drawn with slope 1 and intercept of 0 to indicate how the best-fit line is offset by 1.63 ppm.

A best fit line (solid line) drawn through the scatterplot shows that the ^{13}CA shifts have been systematically shifted upfield by 1.6 ppm, relative to their properly referenced values (dashed line). The same kind of plot could be generated for ^{1}H and ^{15}N chemical shifts to calculate their offset values too. Simple scatterplots such as these could certainly help NMR spectroscopists identify and correct referencing errors -- as long as

an appropriate 3D structure is available. Alternately, by simply calculating the difference between the observed shifts and SHIFTX-calculated shifts for a given nucleus and then averaging these differences, it is also possible to determine the required chemical shift offset. This is precisely what is done in the reference correction program SHIFTCOR (Zhang et al., 2003), available at http://redpoll.pharmacy.ualberta.ca.

Another obvious application of SHIFTX is in chemical shift assignment checking or assignment validation. As anyone who has assigned spectra from larger proteins knows, there is always some uncertainty in the correctness of the chemical shift assignments. Many of these are corrected during the course of subsequent structure determination steps, but some (particularly ^{13}C and ^{15}N shift errors) may not be so easily detected. Shown in Table 3-11 is a list of chemical shift assignments for a small thioredoxin-like protein from Methanobaterium thermoautotrophicum, (Mt0807) which is being studied in our laboratory. Listed in this table are three sets of assignments for ^{15}N, ^{13}CA, ^{13}CB and ^{13}CO shifts for two regions (the N terminus, near the active site, and the C terminus). The first set corresponds to the initial set of assignments which was used in the subsequent structure generation. The second set corresponds to SHIFTX-calculated shifts generated from the initial Mt0807 structure. The last set of shifts corresponds to the corrected, final set of shifts made after a series of manual and automated checks and corrections. Highlighted in bold are the changes that were made, partly as a consequence of the shift predictions generated by SHIFTX. Inspection of this table will indicate that SHIFTX was able to flag a number of residues with large discrepancies between observed and predicted shift values. Furthermore, in

many cases the final "corrected" shifts ended up being much closer to the SHIFTX-calculated shifts. In this way SHIFTX was able to guide the assignment process and clarify a number of assignment ambiguities. Certainly, if a previously existing homologous structure or even an exact X-ray structure is available, it should be possible to use SHIFTX predictions in a similar manner to help guide even the initial assignment process.

Seq	^{15}N	^{15}N	^{15}N	^{13}CA	^{13}CA	^{13}CA	^{13}CB	^{13}CB	^{13}CB	^{13}CO	^{13}CO	^{13}CO
	Old	Shiftx	New	Old	Shift x	New	Old	Shift x	New	Old	Shiftx	New
M 11	115.9	122.0	**121.3**	55.9	56.2	55.9	33.0	32.6	33.0	177.1	175.3	**176.1**
V 12	120.9	127.3	120.9	61.3	62.0	61.3	34.3	28.0	34.3	175.5	173.7	175.5
V 13	126.3	119.3	126.3	62.4	61.4	62.4	32.8	34.4	32.8	175.3	175.3	175.3
N 14	127.0	126.9	127.0	53.9	50.9	53.9	40.5	41.8	40.5	174.7	174.5	174.7
I 15	127.1	129.1	127.1	59.6	59.4	59.6	39.3	41.3	39.3	174.4	174.5	175.4
E 16	127.4	126.4	127.4	54.3	54.4	54.3	34.0	31.4	34.0	176.1	174.9	176.1
V 17	122.4	120.8	122.4	59.6	59.5	59.6	33.9	33.4	33.9	174.4	174.5	174.4
F 18	126.7	128.3	126.7	57.3	56.4	57.3	45.5	42.3	**40.9**	174.7	175.2	**175.5**
T 19	113.6	120.6	**110.9**	57.1	63.4	**59.6**	65.1	72.4	**68..5**	175.6	172.5	**176.7**
S 20	122.4	114.6	**113.6**	59.9	56.9	**56.8**	65.1	64.0	**65.3**	175.0	171.0	**172.0**
P 21	--	--	--	--	63.3	--	--	30.8	--	--	176.5	--
T 22	118.2	115.6	**115.1**	58.4	63.4	58.4	63.9	69.3	63.9	177.6	173.9	**174.7**
C 23	116.3	117.7	**124.8**	57.0	60.7	57.0	30.0	30.7	30.0	174.8	172.8	**177.1**
P 24	--	--	--	--	60.6	57.3	--	32.4	--	--	175.9	--
Y 25	112.0	122.3	112.0	59.6	61.9	59.6	40.9	37.0	40.9	174.2	174.4	174.2
C 26	126.7	114.8	**116.7**	57.3	60.1	**56.1**	45.5	29.6	**30.0**	174.7	171.7	**177.9**

S 82	115.9	115.1	**114.0**	58.4	58.7	**57.1**	63.9	62.0	**65.1**	176.3	175.2	**175.6**
R 83	119.7	119.1	**122.4**	55.4	59.2	**59.8**	29.2	28.8	**29.8**	179.0	178.6	**177.9**
E 84	121.3	118.8	**115.9**	55.7	59.2	**60.6**	33.0	29.2	**28.8**	178.9	178.5	**177.1**
E 85	118.6	117.1	118.6	58.9	59.9	58.9	30.0	29.2	30.0	179.6	179.6	179.6
L 86	120.5	117.3	120.5	57.9	57.8	57.9	40.9	41.3	40.9	179.8	178.9	179.8
F 87	119.8	120.1	119.8	60.5	60.9	60.5	36.7	38.4	36.7	180.4	176.8	180.4
E 88	119.0	120.8	119.0	59.8	58.4	59.8	29.5	29.7	29.5	177.1	179.0	178.7
A 89	120.1	121.8	120.1	54.9	54.9	54.9	18.3	17.9	18.3	177.8	179.6	177.8
I 90	117.8	119.0	117.8	66.2	64.4	66.2	37.9	37.0	37.9	180.4	178.0	180.4
N 91	112.0	117.0	**117.8**	57.8	55.6	**56.9**	37.4	38.6	**38.8**	178.2	176.3	**178.2**
D 92	120.9	119.4	120.9	51.8	56.4	51.8	42.7	40.9	42.7	177.2	178.1	177.2
E 93	124.0	120.3	124.0	59.9	58.6	59.9	32.6	29.3	32.5	176.6	178.6	**178.5**
M 94	116.6	121.3	**125.1**	56.1	57.9	**57.1**	30.4	33.0	**29.9**	177.9	176.4	**177.0**
E 95	118.2	122.4	118.2	56.0	56.7	56.0	33.9	30.4	33.9	176.9	176.2	176.9

Table 3-11. Comparison between initial assignments (Old), SHIFTX calculated assignments (Shiftx) and final assignments (New) for Mt0807. Sequence numbering begins from the leader peptide tag. Assignment changes precipitated by SHIFTX are highlighted in **BOLD**.

As a final example of an application of SHIFTX, we wish to show how chemical shift calculations can be used to evaluate the relative quality of X-ray and NMR protein structures. As has been pointed out by a number of authors (Williamson et al., 1995; Kuszewski et al., 1995a; Laskowski et al., 1996), even the best NMR structures do not appear to have the same quality or effective resolution as most X-ray structures. While the addition of more global restraints (i.e. residual dipolar couplings) or more precisely measured constraints (J-couplings) is certainly helping the situation, there is still a long way to go. The discrepancy between the quality of NMR structures versus X-ray structures can be made particularly evident if we plot out the correlation coefficient between SHIFTX-predicted and observed chemical shifts for ^{1}HA, ^{1}HN, ^{15}N and ^{13}CA nuclei for X-ray structures of varying resolution. These plots, which cover 123 (^{13}CA) to 157 (^{1}HN) proteins each and which exclude most proteins from the SHIFTX training set, are illustrated in Figure 3-4.

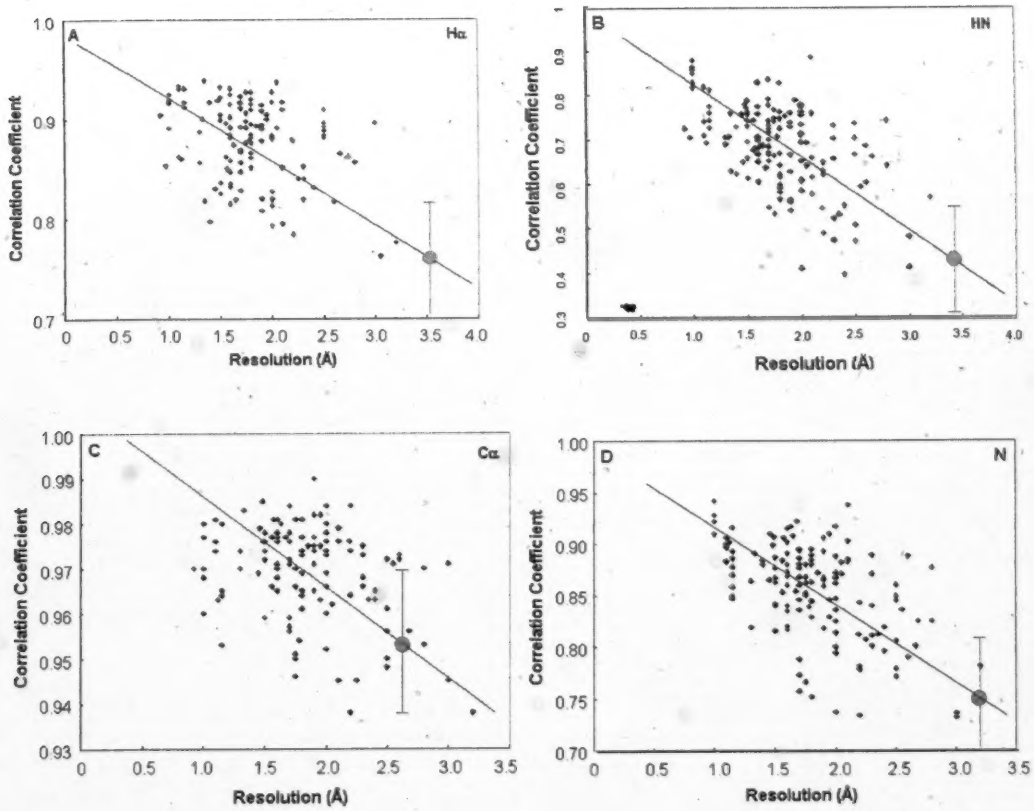


Figure 3-4: Scatterplots of the resolution vs. correlation coefficient for the backbone nuclei corresponding to A) ^{1}Hα; B) ^{1}HN; C) ^{13}Cα; and D) ^{15}N. A best-fit line has been drawn through each graph. The large dot (light gray) with the error bars shown in each graph indicates the position that an average NMR structure would sit on this best-fit line.

As can be seen in all four plots, there is a modest, but obvious trend ($r \sim 0.6$) showing that the agreement between SHIFTX-predicted and observed shifts falls with decreasing resolution. A best-fit trend line is drawn through each of the four distributions. Now if we take all the NMR structures (ranging from 149 structures with ^{13}CA shifts to 229 structures with ^{1}HA shifts) and calculate their average correlation coefficient for each of the four nuclei we find that, on average, NMR structures exhibit relatively poor agreement between SHIFTX-calculated and observed chemical shifts. In fact, if we place a large dot (corresponding to the respective average NMR

correlation coefficients) on the trend lines shown in Figure 4, we can extrapolate that the average protein NMR structure is equivalent to an X-ray structure of 3.0 to 3.5 Å resolution! This resolution estimate is slightly worse than what has been suggested by other authors using different approaches (Williamson et al., 1995; Doreleijers et al., 1998; Laskowski et al., 1996), but given that we are effectively using four independent shift measures, our conclusion appears to be sound.

Interestingly, of all the shifts studied (including, side chain ^{1}H, ^{13}CB, and ^{13}CO) the best indicators of structure quality appear to be ^{1}HN and ^{15}N shifts. These chemical shifts are exquisitely sensitive to H-bonds, aromatic rings, nearby charges and side chain torsion angles (Wishart and Case, 2001). Typically, such subtle structural parameters are not easily captured or measured by traditional NOE measurements, whereas they are more identifiable or at least refineable through higher resolution (<2.0 Å) X-ray crystallography.

What these data help illustrate is the enormous potential that chemical shifts could have in structure refinement. If NMR structures could be fully refined using at least some or even all of their ^{1}H, ^{13}C and ^{15}N chemical shifts, then as these graphs suggest, it does not seem unreasonable to expect that NMR structures could one day match the highest quality X-ray structures -- both in terms of their effective resolution and in terms of their structural "correctness".

Limitations

SHIFTX is not yet capable of predicting all shifts for all nuclei in all proteins. For instance, some rarely measured ^{1}H shifts are not particularly well predicted (HG12 and HE21). This result is most likely due to poor statistical data (needed to model the hypersurfaces) or to atom mislabelling. Likewise SHIFTX does not predict side chain (i.e. beyond CB) ^{13}C and ^{15}N shifts. However, these shifts are rarely measured or reported. Furthermore, they tend not to differ substantively from the random coil values reported previously (Wishart et al., 1995a). A more serious limitation for SHIFTX, however, is the fact that the program does not calculate paramagnetic effects, nor does it account for the presence of organic ligands (heme rings, aromatic substrates, etc.). Parameters and formulae do exist to account for many of these effects (particularly for ^{1}H shifts) for common ligands or metals (Osapay and Case, 1991; Banci et al., 1997; Wishart and Case, 2001) and efforts are underway to incorporate these into the next release of SHIFTX. The inclusion of rare or unique organic ligands (i.e. drug leads, specially developed inhibitors, etc.) will present some challenges and so the parameterization of their effects will likely only be crudely approximated.

Unlike QM approaches, SHIFTX is not particularly sensitive to bond lengths, non-torsional bond angles, bond hybridization state or partial charge distribution. This is both good and bad. At one level, by ignoring these difficult-to-observe effects, it is possible to take unmodified PDB coordinate files and use SHIFTX to predict chemical shifts quite accurately. On the other hand, chemical shifts are exquisitely sensitive to very small coordinate errors and as other workers have shown (Pearson et al., 1997; Le

et al., 1995), when protein structures are "regularized" or minimized to yield optimal covalent geometry, the agreement between QM calculated shifts and observed chemical shifts is substantially improved. While we have shown that SHIFTX is also quite sensitive to the quality of protein structures, this is most likely reflects the quality or precision of non-covalent effects (ring placement, H-bond lengths, torsion angles) rather than in covalent effects such as bond lengths and bond angles. This suggests that the use of QM methods such as those of Xu and Case (2001) or Le et al. (1995) could be of greater assistance in the very detailed refinement of protein structures.

In its present form SHIFTX is not capable of including chemical shift corrections arising from temperature effects, solvent pH effects, local variations in side chain pKa values or isotope effects. Temperature effects are most significant for amide protons and are frequently used to assess H-bond status (Baxter et al., 1997). Temperature does not appear to have a significant effect on other ^{1}H, ^{13}C and ^{15}N shifts. Efforts are underway to include a simple temperature correction term for amide protons in the next release of SHIFTX. In addition to temperature, variations in solvent pH can play a significant role in the quality of chemical shift predictions. This is particularly true for histidine and somewhat less so for other charged amino acids (aspartic acid, glutamic acid, lysine and arginine). It is likely that solvent pH affects the shifts of serine, threonine, cysteine and tyrosine as well. One barrier in modeling these effects has been the difficulty in obtaining reliable solvent pH values for both NMR and X-ray samples. Frequently NMR protein samples are assigned over a range of different pH's, solvents (H_2O, D_2O, DMSO-water) or temperatures and the reported shifts represent either an average shift or a set of heterogeneous shifts collected under different

conditions. This makes it difficult to accurately discern clear pH trends in chemical shifts. On the other hand, for X-ray samples, the true pH of a protein crystal is often difficult to ascertain and is rarely reported in the PDB file. Additionally, the differences between X-ray structures (solved at one pH) and NMR assignments (collected at another pH) further complicate the situation. Modeling solvent pH effects is made even more difficult by the fact that amino acids in proteins will frequently have substantially different pKa values (and titration curves) than free amino acids or unstructured peptides. Given that the prediction of side chain pKa's in proteins is still a difficult computational problem (Gibas and Subramanian, 1996) and given the previously mentioned difficulties in ascertaining accurate pH values, we have chosen to ignore pH effects in this version of SHIFTX. Overall, the inclusion of pH and pKa effects in SHIFTX will require substantially more software development and careful re-measurement of many previously reported shifts under more defined conditions.

Isotope effects can and do affect ^{1}H, ^{13}C and ^{15}N chemical shifts in proteins. Currently SHIFTX does not include deuterium isotope effects in predicting ^{13}C and ^{15}N shifts (Bjorndahl et al., 2001; Gardner et al., 1997). These corrections (about 0.43 ppm for ^{13}CA, 0.82 ppm for ^{13}CB and 0.23 ppm for ^{15}N) will soon be added as an option to the SHIFTX server.

Another limitation of SHIFTX arises from its use of pre-defined sequence and structure parameters. These sequence/structure parameters were originally chosen in 1998 based on their previously reported correlations to chemical shifts (Wishart and Nip, 1998; Osapay and Case, 1991; de Dios et al., 1993). However, since that time, additional properties – including CO and CN bond lengths, backbone omega values,

intrapeptide NH to CO bond distances, etc. have been found to have some impact on measured chemical shifts (Xu and Case, 2001). Hopefully the inclusion of these structural parameters and their corresponding hypersurfaces in future releases of SHIFTX should improve its overall performance.

The fact that SHIFTX uses a hybrid approach drawing on closed form analytical expressions in conjunction with empirical hypersurfaces and look-up tables makes the method somewhat less elegant than QM methods or pure "classical" approaches. It also makes SHIFTX less amenable to incorporation into more standard structure refinement packages, such as AMBER or XPLOR, which rely on having smoothly differentiable functions for conjugate gradient or Newton Raphson minimization. However, we have found that chemical shift optimization using SHIFTX can be done relatively easily and effectively using a Simplex minimizer (which does not require derivatives) or through Genetic algorithms or Monte Carlo searches in torsion angle space. A report on this work will be forthcoming shortly.

Availability

SHIFTX is available as a webserver at http://redpoll.pharmacy.ualberta.ca. A screen shot of the web server is shown in Figure 5. Unsupported copies of the SHIFTX binary code (written in C, compiled on Linux, Solaris, Irix, and Win32) may be obtained on request from the authors.

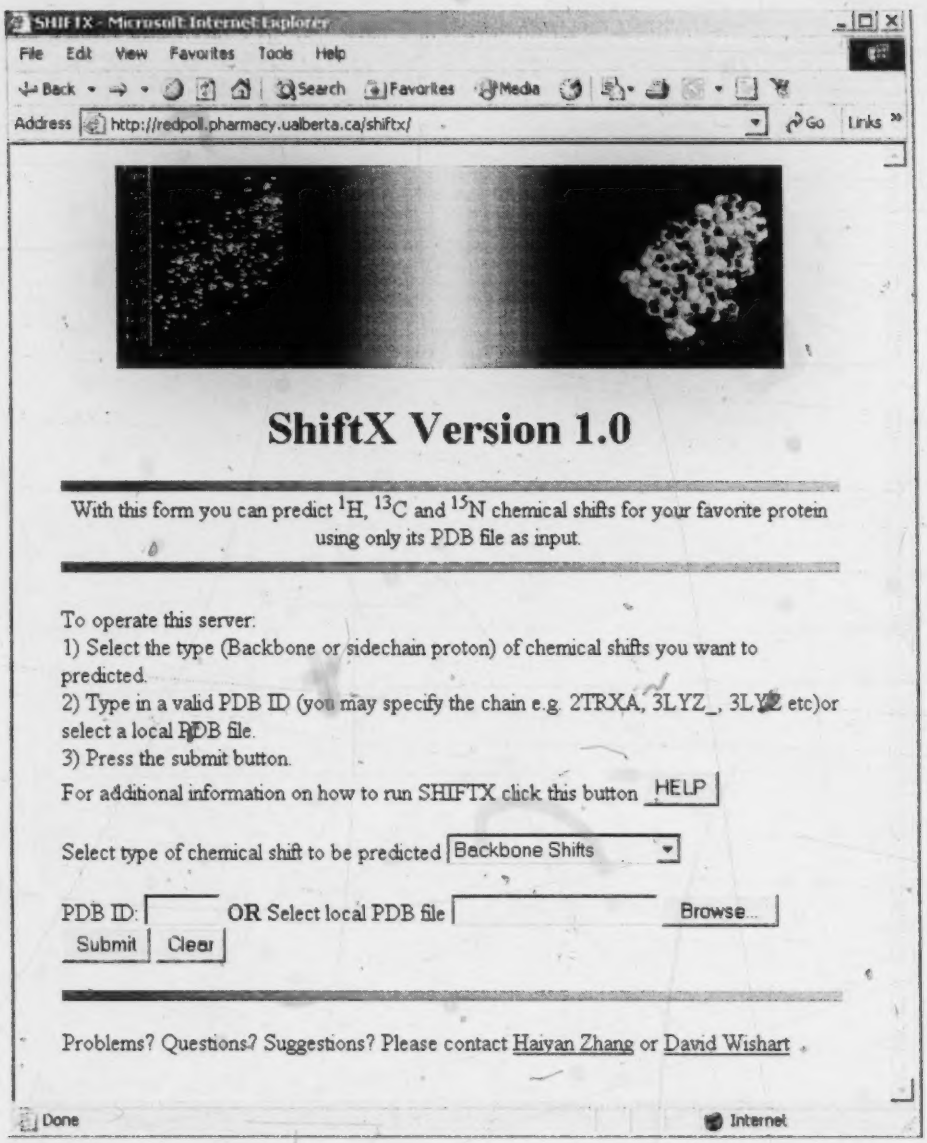

SHIFTX - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://redpoll.pharmacy.ualberta.ca/shiftx/
Go Links
ShiftX Version 1.0
With this form you can predict ^{1}H, ^{13}C and ^{15}N chemical shifts for your favorite protein using only its PDB file as input.
To operate this server:
1) Select the type (Backbone or sidechain proton) of chemical shifts you want to predicted.
2) Type in a valid PDB ID (you may specify the chain e.g. 2TRXA, 3LYZ_, 3LYZ etc)or select a local PDB file.
3) Press the submit button.
For additional information on how to run SHIFTX click this button HELP
Select type of chemical shift to be predicted Backbone Shifts
PDB ID: OR Select local PDB file Browse...
Submit Clear
Problems? Questions? Suggestions? Please contact Haiyan Zhang or David Wishart
Done
Internet

Figure 3-5: A screen shot of the SHIFTX web server.

References

Banci, L., Bertini, I., Savellini, G.G., Romagnoli, A., Turano, P., Cremonini, M.A., Luchinat, C., Gray, H.B. (1997) *Proteins: Struct. Funct. and Genetics*, **29**, 68-76.

Baker, E.N. and Hubbard, R.E. (1984) *Prog. Biophys. Mol. Biol.* **44**, 97-179.

Baxter, N.J. and Williamson, M.P. (1997) *J. Biomol. NMR*, **9**, 359-369.

Beger, R.D. and Bolton, P.H. (1997) *J. Biomol. NMR*, **10**, 129-142.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Research*, **28**, 235-242.

Bjorndahl, T.C., Watson, M.S., Slupsky, C.M., Spyracopolous, L., Sykes, B.D. and Wishart, D.S. (2001) *J. Biomol. NMR*, **19**, 187-188.

Braun, D., Wider, G. and Wuthrich, K. (1994) *J. Am. Chem. Soc.*, **116**, 8466-8469.

Buckingham, A.D. (1960) *Canadian Journal of Chemistry*, **38**, 300 – 307

Case, D.A. (1998) *Curr. Opin. Struct. Biol.*, **8**, 624-630.

Case, D.A. (2000) *Curr. Opin. Struct. Biol.*, **10**, 197-203.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289-302

Dalgarno, D.C., Levine, B.A. and Williams, R.J.P. (1983) *Biosci. Rep.*, **3**, 443-452.

de Dios, A.C., Pearson, J.G. and Oldfield, E. (1993) *Science*, **260**, 1491-1496.

Derewenda, Z.S., Lee, L. and Derewenda, U. (1995) *J. Mol. Biol.* **252**, 248-262.

Doreleijers, J.F., Rullmann, J.A., Kaptein, R. (1998) *J. Mol. Biol.*, **281**, 149-164.

Gardner, K.H., Rosen, M.K. and Kay, L.E. (1997) *Biochemistry*, **36**, 1389-1401.

Gibas, C.J. and Subramanian, S. (1996) *Biophys. J.* **71**, 130-147.

Haigh, C.W. and Mallion, R.B. (1980) *Progress In NMR Spectroscopy*, **13**, 303-344.

Herranz, J., Gonzalez, C., Rico, M., Nieto, J.L., Santoro, J., Jimenez, M.A., Bruix, M., Neira, J.L. and Blanco, F.J. (1992) *Magn. Reson. Chem.*, **30**, 1012-1018.

Iwadate, M., Asakura, T. and Williamson, M.P. (1999) *J. Biomol. NMR*, **13**, 199-211.

Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577-2637.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1995b) *J. Magn. Reson. B.*, **107**, 293-297.

Kuszewski, J., Qin, J., Gronenborn, A.M. and Clore, G.M. (1995a) *J. Magn. Reson. B.*, **106**, 92-96.

Le, H. and Oldfield, E. (1994) *J. Biomol. NMR*, **4**, 341-348.

Le, H., Pearson, J.G., de Dios, A.C. and Oldfield, E. (1995). *J. Am. Chem. Soc.* **117**, 3800-3807.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., Thornton, J.M. (1996) *J. Biomol. NMR.*, **8**, 477-486.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wuthrich, K. (1998) *J. Biomol. NMR*, **12**, 1-23.

Osapay, K. and Case, D.A. (1991) *J. Am. Chem. Soc.*, **113**, 9436-9444.

Osapay, K. and Case, D.A. (1994) *J. Biomol. NMR*, **4**, 215-230.

Osapay, K., Theriault, Y., Wright, P.E. and Case, D.A. (1994). *J. Mol. Biol.*, **244**, 183-197.

Pearson, J.T., Le, H., Sanders, L.K., Godbout, N., Havlin, R.H. and Oldfield, E. (1997) *J. Am. Chem. Soc.* **119**, 11941-11950.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217-236.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490-5492.

Wagner, G., Pardi, A. and Wuthrich, K. (1983) *J. Am. Chem. Soc.*, **105**, 5948-49.

Williamson, M. P. and Asakura, T. (1997) *Methods Mol. Biol.*, **60**, 53-69.

Williamson, M.P., Asakura, T., Nakamura, E. and Demura, M. (1992) *J. Biomol. NMR*, **2**, 93-98.

Williamson, M.P., Kikuchi, J. and Asakura, T. (1995) *J. Mol. Biol.*, **247**, 541-546.

Wishart, D.S. and Case, D.A. (2001) *Methods Enzymol.*, **338**, 3-34.

Wishart, D.S. and Nip, A.M. (1998) *Biochemistry & Cell Biology*, **76**, 153-163.

Wishart, D.S. and Sykes, B.D. (1994) *Methods Enzymol.*, **239**, 363-392.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995a) *J. Biomol. NMR*, **5**, 67-81.

Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995b) *J. Biomol. NMR*, **6**, 135-40.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1991) *J. Mol, Biol.*, **222**, 311-333.

Wishart, D.S., Willard, L., Richards, F.M., and Sykes, B.D. (1994) "VADAR: A comprehensive program for protein structure evaluation. Version 1.2." Edmonton, Alberta, Canada.

Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) *J. Mol. Biol.*, **285**, 1733-1747.

Xu, X-P. and Case, D.A. (2001) *J. Biomol. NMR*, **21**, 321-333.

Xu, X-P. and Case, D.A. (2002) *Biopolymers*, **65**, 408-423.

Zhang, H., Neal, S., and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173-195

Chapter 4: Conclusion

The research described herein required the creation of two software tools: one, MINER, for discerning and extracting patterns from large data sets, and the other, SHIFTX, using those patterns along with a variety of other equations to predict chemical shifts from protein structures.

The second chapter of this document describes the development and function of MINER in detail. The goal of the project was finding useful patterns and trends in a large, complex body of data. This extremely general goal as stated is a task of unlimited scope, and so the first part of the project was defining the limits of the search. The most far-reaching limitation was to restrict MINER to searching for one and two-dimensional tables that correlated one or two protein attributes with chemical shifts. Other important initial choices were the list of the protein attributes used in the input and the division of continuously-valued attributes into appropriately discretized bins. Most important was the choice of objective criteria used to evaluate proposed tables: the correlation coefficient and the RMSD between the predicted and actual chemical shift values.

The input to MINER was a collection of high resolution (less than or equal to 2.0 Angstrom) protein structures derived from X-ray data, along with experimentally determined chemical shifts for those structures. From this data was derived a database with each record representing an individual amino acid and its attributes, as well as its experimental chemical shift, and a set of precomputed chemical shift contributions from well-characterized phenomena such as shifts due to ring currents and electrostatic effects. These precomputed values were summed to generate an initial "best guess" of the chemical shift. MINER would then generate trial tables from this database and select

those which did the best job of minimizing the difference between the "best guess" predicted chemical shift and the experimental value. The best table of each iteration would be chosen, and its effects added to the best guess. This search procedure would be repeated until no further improvement in quality (RMSD and/or correlation) was seen. In order to improve the quality, reliability, and generality of MINER's predictions, several mathematical and statistical techniques were tried, not all successfully.

Chapter 3 describes the theory behind and development of SHIFTX, which applies classical and semi-classical equations and statistically-derived tables from MINER to quickly and accurate predict chemical shifts for any protein structure. SHIFTX ties together the findings of a number of researchers in the field of protein NMR, as well as the body of classical and empirical equations predicting the effects of various well-understood physical phenomena.

SHIFTX reads protein structures in the standard PDB file format as a starting point for its predictions. Any missing hydrogen atoms are filled in using standard heuristics, and the resulting structure is used to compute the effects of hydrogen bonding, ring currents, and electric fields. The MINER-derived tables are then applied, and the sum of all these factors are reported as the predicted chemical shifts.

The accuracy of SHIFTX predictions was validated simply by comparing predicted with actual chemical shifts by computing the correlation and RMSD values between the two sets of shift. This validation was performed for both the "training" data that was used to calibrate SHIFTX, and "test" data which the program had not seen during its development. Instances of significant disagreement between SHIFTX and experimental data were often found to be the result of misassignments in the NMR data

or incorrect structures. One important application for SHIFTX will be to validate new NMR chemical assignments, where significant disagreements between theory and experiment will point out areas that need to be carefully examined. SHIFTX has also found application in checking and, when necessary, correcting chemical shift referencing, which has suffered from a lack of standardization in the past. SHIFTX was also used to explore the implications of structural resolution in NMR-derived protein structures, and quantify the relative accuracy NMR and x-ray structures.

There are two primary directions for future work in correlating chemical shifts with protein structures. The first, most obvious one is improving the accuracy of predictions. As suggested in the MINER chapter, one way to go about this would be increasing the number of factors that MINER uses to build its tables to three or more. The key limitation here is the amount of time required to search the space of high-dimension tables; the time required for an exhaustive search increases exponentially with the number of factors in the table. Adaptations of this exhaustive strategy, such as finding two-dimensional tables that work well and augmenting them with additional factors, may prove more practical. An increase in "lookahead", where MINER simultaneous constructs and tests two or more tables (applied in sequence) might also yield more accurate results, and also increase the rate at which MINER converges on a correct solution (assuming the use of feedback/optimization). MINER could also benefit greatly from careful coding optimizations that take advantage of processor-level data caches and new instruction set enhancements that offer SIMD floating point operations (Intel, 2003).

The predictive accuracy of SHIFTX could benefit from improvements in calibrations and refinement of empirically-derived constants in particular, such

improvements stemming from the availability of new structures and chemical shift data, and more particularly from the increased knowledge of how various factors influence chemical shifts that has come from MINER research.

The other future direction to consider is taking what has been learned from SHIFTX research and apply it in the opposite sense: predicting protein structure from chemical shifts. This remapping is not a simple one: each combination of physical factors can only correspond to one chemical shift, but a given chemical shift can correspond to many different ranges of factors. Even in its present state, SHIFTX can be used to refine an existing structure by applying Monte Carlo (Heermann, 1990) or Genetic Algorithm (Koza, 1992; Deaven and Ho, 1996) methods towards reducing the discrepancy between predicted and experimental shifts. A more advanced application of this same principle would be to start with nothing but a protein sequence and the chemical shifts of a few of its nuclei, and use technology from MINER and SHIFTX to determine the backbone torsion angles. Once these have been determined to within a reasonable range, more tables and formulas could be applied to refine the "first draft" structure, this latter iteration beginning to determine chi angles, the presence of hydrogen bonds, and other key structural elements. A reasonable first step toward this goal would be a study of how large an area of the Ramachandran plot (Ramachandran and Sasisekharan, 1968) corresponds to the chemical shift of a given single nucleus, and thence to chemical shifts from multiple nuclei; for example, the alpha carbon shift alone might only eliminate 10% of the allowed Ramachandran area, but the alpha carbon and nitrogen shifts together might point to a significantly smaller region. This work of applied statistics would be of

significant value in determining just how much chemical shift information is necessary to determine structural parameters with useful accuracy.

In conclusion, this project has demonstrated that there are predictable and quantifiable relationships between the structure of proteins and their chemical shifts that can be modeled accurately without significant computational resources, and that these relationships are sufficiently well defined to be useful in practical applications. Armed with this knowledge, it may be possible to greatly accelerate the process of determining and refining protein structures, firstly by allowing the use of easily-acquired one-dimensional chemical shifts to determine structure, and secondly by permitting significant automation of the process. With the solution of protein structures being a significant limiting step in utilizing the information generated by efforts such as the Human Genome Project, the prospect of simplifying and speeding up the process justifies further research in this area.

References

Deaven, D. and Ho, K. (1995) *Phys. Rev. Lett.*, **75**, 288-291

Heermann, D.W. (1996) *Computer Simulation Methods in Theoretical Physics*, Springer-Verlag, New York, NY

Intel Corporation (2003) *IA-32 Intel Architecture Software Developer's Manual Volume 2: Instruction Set Reference*. Intel Corporation, Denver, CO.

Koza, J.R. (1992) *Genetic Programming*, MIT Press, Cambridge, MA

Ramachandran, G.N. and Sasisekharan, V. (1968) *Adv. Protein Chem.*, **23**, 283-438